Codon usage bias in Archaea

Laura R. Emery

Ph.D.

University of Edinburgh

2010

for Mom

ACKNOWLEDGEMENTS

P.M. Sharp has provided me with this superb opportunity to pursue an area of research which has thoroughly intrigued me. I thank him for his experience, expertise and unsurpassed attention to detail.

I thank the numerous members of I.E.B for their guidance and support throughout my time in Edinburgh. A particular thanks to B.C. and D.C. Charlesworth who have made me feel very welcome at their lab group, where I have learned a lot. The recent computational support from R.H.J. Perry has been invaluable and I thank P.R. Haddrill, J.J. Welsh, K. Zeng and S.J. Lycett for advice and/or discussion.

I thank my school teachers for inspiration and C.M. Wade for the advice to embark upon this project in Edinburgh. The encouragement and endurance from my friends, J. Raghwani and M.J. Ward, has been greatly appreciated.

I thank my family: B.A. Evans, S.C. Emery and M.W.H. Emery for their infallible support and tolerance. I would thank H.C. Emery if it were possible; she always insisted that I pursue my dreams, and yet ironically, has exemplified the roles of mutation and natural selection in bringing life in and out of existence.

DECLARATION

The work of this thesis is my own unless stated otherwise.

Laura R. Emery

2010

PUBLICATIONS

SHARP, P. M., L. R. EMERY and K. ZENG, 2010 Forces that influence the evolution of codon bias. Phil. Trans. R. Soc. **365**: 1203-1212.

EMERY, L. R., and P. M. SHARP, 2011 Impact of translational selection on codon usage bias in the archaeon *Methanococcus maripaludis*. Biol. Lett. doi: 10.1098/rsbl.2010.0620.

CONTENTS

7. THE IDENTITY AND DIVERGENCE OF OPTIMAL CODONS IN ARCHAEA

8. EXPLORING THE STRENGTH OF SELECTED CODON USAGE BIAS FOR ACCURATE TRANSLATION IN ARCHAEA

9. CONCLUSIONS

ABSTRACT

Synonymous codon usage bias has been extensively studied in Bacteria and Eukaryotes and yet there has been little investigation in the third domain of life, the Archaea. In this thesis I therefore examine the coding sequences of nearly 70 species of Archaea to explore patterns of codon bias. Heterogeneity in codon usage among genes was initially explored for a single species, *Methanococcus maripaludis*, where patterns were explained by a single major trend associated with expression level and attributed to natural selection. Unlike the bacterium *Escherichia coli*, selection was largely restricted to two-fold degenerate sites.

Analyses of patterns of codon usage bias within genomes were extended to the other species of Archaea, where variation was more commonly explained by heterogeneity in G+C content and asymmetric base composition. By comparison with bacterial genomes, far fewer trends were found to be associated with expression level, implying a reduced prevalence of translational selection among Archaea. The strength of selected codon usage bias (*S*) was estimated for 67 species of Archaea, and revealed that natural selection has had less impact in shaping patterns of codon usage across Archaea than across many species of Bacteria. Variation in *S* was explained by the combined effects of growth rate and optimal growth temperature, with species growing at high temperatures exhibiting weaker than expected selection given growth rate. Such a relationship is expected if temperature kinetically modulates growth rate via its impact upon translation elongation, since rapid elongation rates at high temperatures reduce the selective benefit of optimal codon usage for the efficiency of translation. Consistent with this, growth temperature is negatively correlated with minimal generation time, and numbers of rRNA operons and tRNA genes are reduced at high growth temperatures. The large fraction of thermophilic Archaea relative to Bacteria account for the lower values of *S* observed.

Two major trends were found to describe variation in codon usage among archaeal genomes; the first was attributed to GC3s and the second was associated with arginine codon usage and was linked both with growth temperature and the genome-wide excess of G over C content. The latter is unlikely to reflect thermophilic adaptation since the codon primarily underlying the trend appears to be selectively disfavoured. No correlations were observed with genome wide GC3s and optimal growth temperature and neither was GC3s associated with aerobiosis.

The identities of optimal codons were explored and found to be invariant across U and C-ending two-fold degenerate amino acid groups. The identity of optimal codons and anticodons across four and six-fold degenerate amino acid groups was found to vary with mutational bias. As was first observed in *M. maripaludis*, selected codon usage bias was consistently greater across two-fold relative to four-fold degenerate amino acid groups across Archaea. This broad pattern could reflect ancestral patterns of optimal codon divergence, prevalent among four-fold but not two-fold degenerate amino acid groups. Consistent with this, the strength of selected codon usage bias was found to be reduced following the divergence of optimal codons, and implies that optimal codon divergence typically proceeds following the relaxation of selection.

Finally, a method was developed to partition the strength of selection (S) into separate components reflecting selection for translational efficiency (S_{eff}) and selection for translational accuracy (S_{acc}) by comparing the codon usage across conserved and non-conserved amino acid residues. While estimates of S_{acc} are somewhat sensitive to the designation of conserved sites, a general pattern emerged whereby accuracy-selected codon usage bias was consistently strongest across a subset of the most highly conserved sites. Several estimates of S_{acc} were consistently higher than the 95% range of null values regardless of the dataset, providing evidence for accuracy-selected codon usage bias in these species.

1. INTRODUCTION

The genetic code is inherently redundant with 64 codons, the triplet combinations of four nucleotide bases, encoding 20 amino acids. Synonymous codons are those which encode the same amino acid and typically vary at the third position. During protein synthesis codons are translated following the initial base pairing of cognate tRNA anticodons at the ribosomal A site. The tRNA specifies the amino acid to be incorporated into the polypeptide chain. In most species, there are between 31 and 46 different tRNA anticodons meaning that certain tRNAs recognise more than one codon. This is largely achieved through 'wobble' base pairing (CRICK 1966), whereby the first tRNA anticodon position is less spatially confined to permit non-standard base pairing and thus the recognition of multiple third codon positions. Chemical modification at the first anticodon position can act to improve the efficiency of wobble base pairing (HOLLEY *et al.* 1965; VARANI and MCCLAIN 2000).

The usage of alternative synonymous codons does not impact upon the polypeptide chain sequence and so might be expected to be devoid of phenotypic consequences. Yet in a wide range of organisms alternative synonymous codons are not used at equal frequencies (GRANTHAM *et al.* 1981; SHARP *et al.* 1988). Non-random patterns of codon usage vary between species and among genes within the same species and can be explained as a balance of mutational biases, natural selection and random genetic drift (BULMER 1991a; SHARP and LI 1986). Natural selection favours codons best recognised by the most abundant tRNAs (IKEMURA 1981; IKEMURA 1985), for efficient and accurate translation (ANDERSSON and KURLAND 1990; EHRENBERG and KURLAND 1984).

Much of the early work deducing the processes shaping biases in codon usage focused upon the genetic model systems *Escherichia coli* and *Saccharomyces cerevisiae* (BENNETZEN and HALL 1982; GRANTHAM *et al.* 1981), where tRNA abundances were characterised (IKEMURA 1981; IKEMURA 1982). Codon usage bias rapidly emerged as a common feature among unicellular organisms (SHARP and LI 1986), including the Bacteria *Salmonella enterica* (SHARP and LI 1987b), *Mycoplasma capricolum* (MUTO *et al.* 1985) and *Bacillus subtilis* (SHIELDS and SHARP 1987), as well as the Eukaryotes *Saccharomyces* (SHARP *et al.* 1988), *Dictyostelium discoideum* (SHARP and DEVINE 1989), *Tetrahymena* (MARTINDALE 1989) and *Chlamydomonas* (CAMPBELL and GOWRI 1990). Now, our knowledge of codon usage in Bacteria is extensive, with analyses of large numbers of fully sequenced genomes of broad taxonomic range. Codon

usage has also been well-studied in certain groups of Eukaryotes. Yet Bacteria and Eukaryotes are just two of three domains of life. The third domain is the Archaea, no more closely related to Bacteria than to Eukaryotes (FOX 1980). Despite their discovery 30 years ago, Archaea remain a largely understudied taxonomic group and little is known of their codon usage. In this study I therefore investigate patterns of synonymous codon usage bias among Archaea. Here I review what is known of codon usage in all domains of life, with particular emphasis upon patterns observed among Bacteria. Codon usage in Bacteria is most extensively studied and likely to be of most relevance to Archaea, since both Bacteria and Archaea are exclusively unicellular, largely haploid, and possess similar sized and generally circular chromosomes, with similar mechanisms for gene regulation and control.

1.1 Codon usage bias in Bacteria

1.1.1 Mutational biases

Mutational biases occur when mutational rates among the nucleotide bases are not equal, and appear to be a widespread feature of bacterial genomes, where mutation rates are influenced by numerous varied mechanisms (LOBRY 1996; ROCHA *et al.* 2006). In the absence of selection and at equilibrium, biased patterns of mutation are reflected in base composition (SUOEKA 1962) and it is known that bacterial genomes vary considerably in G+C content (MUTO and OSAWA 1987). Variation in base composition among bacterial genomes is greatest at third codon positions (SHARP *et al.* 2005) where sites are often synonymously variable and subject to less functional constraint; thus mutational biases impact upon codon usage.

1.1.1.1 Variation in mutational patterns among bacterial genomes

The greatest source of variation in codon usage among bacterial genomes has been attributed to heterogeneity in base composition. A variety of multivariate analyses of codon usage among genomes consistently identify a primary trend that is associated with G+C content (CHEN *et al.* 2004; LOBRY and NECSULEA 2006; LYNN *et al.* 2002). Variation in G+C content among genomes is commonly believed to reflect genome-specific mutational patterns, consistent with the observations that as few as 10 genes are required to reliably predict genome G+C content (ZAVALA *et al.* 2005), that 90% of variance in G+C content occurs between rather than within genomes, and that intergenic G+C content can be use to predict codon usage bias (CHEN *et al.* 2004).

1.1.1.2 Selection upon base composition?

Whilst attributing variation in G+C content to mutational biases is the most parsimonious explanation, there has been some speculation that base composition is subject to natural selection. First, it was suggested that G+C nucleotides would be selectively favoured in organisms living at high temperatures due to the increased thermostability of G≡C base pairing (ARGOS *et al.* 1979). Consistent with this view, correlations of genome G+C content and optimal growth temperature were observed within several bacterial families (MUSTO *et al.* 2004) but few are statistically supported and a more comprehensive study of 764 species failed to observe any trend between genome G+C content and optimal growth temperature (GALTIER and LOBRY 1997). Multivariate analyses of codon usage among genomes have identified a trend, distinct and orthogonal to G+C content, and associated with growth temperature (LOBRY and CHESSEL 2003; LOBRY and NECSULEA 2006; LYNN *et al.* 2002). Whilst the trend was initially interpreted as reflecting the adaptation of base composition to high temperatures (LYNN *et al.* 2002), it was found to be largely caused by atypical arginine codon usage, and might simply reflect a mutational pattern associated with growth temperature (LOBRY and NECSULEA 2006).

An association of aerobiosis and G+C content was observed among 195 genera of Bacteria and Archaea; aerobes were found to be an average of 7% more G+C rich than anaerobes (NAYA *et al.* 2002). The association was interpreted as reflecting selection upon amino acid composition since amino acids found to be overrepresented among anaerobes contained 9/14 A+T nucleotides at nonsynonymous sites. Yet 9/14 is not significantly different to that expected from a random distribution ($\chi^2 = 1.1$; $p = 0.30$) and no attempt was made to examine the G+C content of synonymously variable sites (GC3s) in order to attribute the effect solely to non-synonymous residues. The effect could be explained more simply as a mutational bias linked with metabolism.

Correlations of genome size and genome G+C content among bacteria (BASTOLLA *et al.* 2004; MUSTO *et al.* 2006), which are present with the inclusions of small A+T rich endosymbiont genomes, have been interpreted as reflecting less effective purifying selection for G+C bases in these species (BASTOLLA *et al.* 2004; MANN and CHEN 2010). However the low G+C content observed among endosymbionts has been linked with the loss of DNA repair systems (DALE *et al.* 2003; KLASSON and ANDERSSON 2006; LIND and ANDERSSON 2008) and mutational

patterns inferred within and among *Buchnera* species are consistent with current base composition reflecting that expected at mutational equilibrium (WERNEGREEN and FUNK 2004), so for these species there appears to be no requirement for an adaptive explanation.

Finally, recent studies have examined patterns of polymorphism across Bacteria, and find consistent excesses of GC→AT relative to AT→GC mutations in many species (HERSHBERG and PETROV 2010; HILDEBRAND *et al.* 2010), across both coding and non-coding sequences (HERSHBERG and PETROV 2010), with the proportion of GC→AT over AT→GC polymorphisms positively correlated with genome-wide G+C content (HILDEBRAND *et al.* 2010). This pattern is not easily explained by genome-specific mutational biases since that would imply widespread deviations from mutational equilibria. Rather these observations have been interpreted as the action of natural selection, or other non-neutral processes such as biased gene conversion (explained in section 1.2.1). However these interpretations remain difficult to reconcile. Biased gene conversion does not appear to be a likely candidate since excesses of GC→AT polymorphisms were observed in Bacteria lacking recombination where there is no opportunity for gene conversion to take place (HILDEBRAND *et al.* 2010), and it remains a mystery why selection might favour alternative base compositions among species.

1.1.1.3 Variation in mutational patterns within bacterial genomes

Patterns of codon usage vary among genes within bacterial genomes and are commonly identified using multivariate approaches (GREENACRE 2007; KLOSTER and TANG 2008; LOBRY and CHESSEL 2003; SUZUKI *et al.* 2008). The major sources of variation are due to (i) natural selection for translationally optimal codons associated with expression level, (ii) heterogeneity in G+C content associated with recently horizontally transferred genes and (iii) heterogeneity in G+T content associated with strand specific mutational biases.

The acquisition of foreign genes by bacterial species is widespread (OCHMAN *et al.* 2000) and may proceed through a variety of routes, following the uptake of foreign DNA, phage infection and conjugation. Sources of foreign DNA can often be subject to rather different mutational patterns than those of the host chromosome, and in these instances it is often possible to detect lateral transfer events on the basis of abnormal codon usage alone (GARCIA-VALLVE *et al.* 2000; MEDIGUE *et al.* 1991), provided that insufficient time has passed for the acquired genes to ameliorate to their new mutational equilibria (LAWRENCE and

OCHMAN 1997). However there can be problems with this approach (KOSKI *et al.* 2001). The observation that horizontally acquired genes are typically A+T rich (DAUBIN *et al.* 2003; LAWRENCE and OCHMAN 1997) is consistent with a major role for extra-chromosomal elements in genetic exchange, given the relative A+T richness of plasmids compared with their respective host genomes (VAN PASSEL *et al.* 2006). In the genome of the G+C rich gamma-proteobacterium *Pseudomonas aeruginosa*, the primary trend shaping codon bias was found to be associated with foreign genes with abnormally low G+C content (GROCOCK and SHARP 2002). Thus continual acquisition of horizontally transferred genes can impact very strongly upon patterns of codon usage within genomes.

Genes on the leading strand with respect to DNA replication often exhibit alternative base composition to those on the lagging strand (LOBRY 1996; MCLEAN *et al.* 1998). Whilst the magnitude and direction of these compositional differences can vary among species, typically genes on the leading strand are G+T rich (FRANCINO and OCHMAN 1997). In some genomes skew patterns can be very strong, and in the spirochaete *Borrelia burgdorferi,* they explain the greatest fraction of variation in patterns of codon bias among genes (LAFAY *et al.* 1999). Yet in other genomes such as *Synechoscystis* sp. PCC6803 (MCLEAN *et al.* 1998) strand skews can be slight, and so are most conveiently examined by assessment of the cumulative excess of G over C (or T over A) across the chromosome. Cumulative skew analyses typically show an abrupt sign change at the origin and terminus of replication, and are clearest when restricted to the base composition across third codon positions (LOBRY 1996). Therefore strand skews are interpreted as reflecting mutational biases associated with DNA replication. Whilst several studies have attempted to identify the mechanism underlying these patterns (FRANCINO and OCHMAN 2001; KLASSON and ANDERSSON 2006; WORNING *et al.* 2006), the tremendous variability in skew patterns among genomes implies that there are numerous molecular causes (NECSULEA and LOBRY 2007; ROCHA *et al.* 2006). Replication-associated biases can impact upon G+C content (DAUBIN and PERRIERE 2003) but the size of these effects is very small, with one possible exception (KERR *et al.* 1997; MCINERNEY 1997). Other asymmetrical mutational biases have been associated with gene expression level (LOBRY and SUEOKA 2002) although these are generally weaker than those associated with replication (KLASSON and ANDERSSON 2006; NECSULEA and LOBRY 2007). In conclusion, biased patterns of mutation of many kinds are important forces shaping patterns of codon usage within and among bacterial genomes.

1.1.2 Natural selection

Here, studies in Bacteria exemplify early research deducing the nature of selected codon usage bias, but note that many of these findings were also elucidated in yeast.

1.1.2.1 Evidence for selected codon usage bias

Patterns of codon usage bias are often more complex than would be predicted from simple mutational biases. In *Escherichia coli*, codons occurring most frequently among highly expressed genes for the four fold degenerate amino acids glycine and proline take the forms NNU and NNG respectively, see Table 1 in (SHARP and LI 1987a). Since these abundant codons differ in the nucleotide at their third codon position, their high frequency could not be the product of a simple mutational bias. Several observations indicate that natural selection is responsible for these complex patterns of codon bias. Firstly, in a study of 600 gene sequences across 107 species, patterns of synonymous codon usage bias were shown to vary with gene expression levels obtained for *E. coli* (GOUY and GAUTIER 1982). Further intragenomic correlations of gene expression level with measures of codon usage bias have been established for a variety of bacterial species including *E. coli* (IKEMURA 1985), *Streptococcus pneumoniae* (MARTÍN-GALIANO *et al.* 2004) and *Lactobacillus lactis* (DRESSAIRE *et al.* 2009). Such patterns of bias among highly expressed genes are often distinctive, vary among species, and may even oppose the prevailing direction of mutational pressure. For instance among two-fold degenerate sites in the A+T rich bacterium *Clostridium perfringens*, U-ending codons are common across the vast majority genes, and yet it is the C-ending synonyms which dominate the sequences of the highly expressed genes (MUSTO *et al.* 2003). It is difficult to explain these complex patterns as the product of mutational biases alone.

It has been demonstrated that codons overrepresented in highly expressed genes are complementary to the tRNA species with the greatest experimentally determined intracellular abundances, for both *E. coli* (IKEMURA 1981) and *B. subtilis* (KANAYA *et al.* 1999). In *E. coli* for example, of the four tRNAs occurring at experimentally detectable levels for the amino acid leucine, it is the tRNA containing the CAG anticodon which is present at the highest concentrations, and perfectly complements the CUG codon prevalent among highly expressed genes (IKEMURA 1985). The identity of the most abundant tRNAs and complementary codons can vary among species. For instance by contrast with *E. coli*, in *B. subtilis* it is the tRNA with the UAA anticodon which is most abundant for the amino acid

leucine and complements the UUA codon occurring at high frequencies within highly expressed genes (KANAYA *et al.* 1999). The covariance of the codon usage of highly expressed genes with tRNA abundances cannot be explained by neutral processes and implies that both traits coevolve.

Interspecific comparisons of *E. coli* and *S. enterica* have revealed further evidence for the action of natural selection upon synonymous sites. Lower synonymous substitution rates were observed in the highly expressed genes of *E. coli*, consistent with the action of purifying selection (SHARP and LI 1987b), although it seems that codon bias alone cannot account for the magnitude of the observed reduction in divergence (BERG and MARTELIUS 1995 ; EYRE-WALKER and BULMER 1995).

Thus a consistent view has emerged that natural selection acts upon highly expressed genes, favouring those codons best recognised by the most abundant tRNA species (IKEMURA 1985) for accurate and efficient translation. Selectively beneficial synonyms are termed optimal codons and may be putatively identified as those significantly overrepresented among highly expressed genes (HENRY and SHARP 2007). Using this approach to identify the optimal codons of *E. coli* and *C. perfringens*, Sharp and colleagues examined the frequency spectrum of optimal codons across polymorphic sites to assess the evidence for ongoing selection on codon bias (SHARP *et al.* 2010). In the situation that patterns of codon bias evolve neutrally, influenced only by mutational biases and random genetic drift, the site frequency spectrum for optimal codons is expected to follow a U-shaped distribution provided that there is no population structure and assuming free recombination among sites (MCVEAN and CHARLESWORTH 1999). The action of natural selection is expected to skew the distribution towards higher frequencies of optimal codons since selection acts efficiently to prevent non-optimal codon variants from becoming fixed, but is less efficient at eliminating their presence from low frequencies in a population. Frequency distributions of optimal codons in *E. coli* and *C. perfringens* were found to be skewed towards high frequency variants across highly expressed genes, and as such are difficult to explain by the action of neutral processes alone (SHARP *et al.* 2010). So there is consistent evidence, through many lines of enquiry, that the codon usage of highly expressed genes is subject to natural selection.

1.1.2.2 Variation in translational selection among Bacteria

Whilst there is clear evidence that natural selection shapes patterns of codon usage within many bacterial genomes (IKEMURA 1981; MUSTO *et al.* 2003; SHIELDS and SHARP 1987), the extent to which selection impacts upon codon bias varies among species, and for some species it seems that selection is absent altogether (ANDERSSON and SHARP 1996; HERBECK *et al.* 2003; KERR *et al.* 1997; LAFAY *et al.* 1999). One such example is the ulcer-causing bacterium *Helicobacter pylori* where highly expressed genes show no differential patterns of codon bias, do not obviously complement with predicted tRNA anticodons (LAFAY *et al.* 2000), and do not exhibit skewed optimal codon frequency spectra across polymorphic sites (SHARP *et al.* 2010). Several studies have examined large numbers of bacterial genomes to investigate variability in the strength of selection on codon bias among species (DOS REIS *et al.* 2004; ROCHA 2004; SHARP *et al.* 2005; SHARP *et al.* 2010). In a study of 102 species, Rocha and colleagues obtained measures for the degree of codon bias that was due to selection by contrasting ENC (general measure of the degree of codon bias) for highly expressed genes with ENC for all of the genes within a genome. The convenience of this approach is that the ENC statistic does not require the *a priori* identification of optimal codons, however its generality may encapsulate additional forms of bias rather than those specific to translational selection. A different study has estimated the strength of selected codon bias by means of contrasting the codon frequencies of highly expressed genes with those observed genome wide (SHARP *et al.* 2005) under a population genetics model of selection-mutation-drift equilibrium (BULMER 1991b). Optimal codon frequencies were computed for the two-fold degenerate amino acid groups: phenylalanine (Phe), isoleucine (Ile), tyrosine (Tyr) and asparagine (Asn), because the C-ending codon is always optimal for these codon families. With the assumptions that codon frequencies among highly expressed genes reflect those at selection-mutation-drift equilibrium, whilst codon frequencies genome-wide are influenced only by mutational biases and random genetic drift, it was possible to estimate S, the strength of selection, which corresponds to the population genetic parameter $2N_e s$ (where N_e is the effective population size and s is the per site selection coefficient on codon usage).

Consistent trends emerge from these interspecific comparative analyses of selected codon bias. Numbers of rRNA operons within a genome, known to impact upon growth rate (CONDON *et al.* 1995; STEVENSON and SCHMIDT 2004), are negatively correlated with minimal generation time among genomes (SHARP *et al.* 2010; VIEIRA-SILVA and ROCHA 2010), implying

a widespread mechanism for upregulation of ribosome synthesis, via gene copy number, for rapid growth (DETHLEFSEN and SCHMIDT 2007). Similarly, numbers of tRNA genes are highly correlated with respective tRNA abundances (DONG *et al.* 1996; KANAYA *et al.* 1999) and total numbers of tRNA genes per genome were found to be negatively correlated with generation time (ROCHA 2004; SHARP *et al.* 2010). The strength of selected codon usage bias is positively correlated with the numbers of ribosomal RNA operons, the numbers of tRNA genes, and is negatively correlated with generation time (ROCHA 2004; SHARP *et al.* 2005; SHARP *et al.* 2010), indicating that selection is most effective in species with rapid growth rates. Collectively these observations are interpreted as reflecting a conserved strategy, through the coevolution of selected codon bias, rRNA operons and tRNA genes, to achieve rapid growth.

While the direction of selected codon bias appears to be invariant for the C-ending two-fold degenerate families, the identity of optimal codons for other amino acids can vary among species. For instance, the CUG codon is optimal for the amino acid leucine in *E. coli* (IKEMURA 1985) and yet it is the UUA codon which is optimal in *B. subtilis* (KANAYA *et al.* 1999). Since optimal codons are coadapted with their respective tRNA genes, any change in either codon usage or tRNA gene content is expected to incur a fitness cost; so it remains unclear exactly how the identities of optimal codons diverge. A neutral explanation requires there to be weak selected codon usage bias and/or a low effective population size in the ancestral lineage; analogous to the general mechanisms for crossing fitness landscapes first propsed by Wright (WRIGHT 1932). Then, a directional change in mutational bias may gradually impact upon patterns of codon usage bias, such that resumption of a selective regime could result in alternative optimal codon identities. Alternatively, the divergence of optimal codons might be driven by directional mutational pressure rather than genetic drift (SHIELDS 1990). Gradual changes in patterns of mutation can impact upon the codon usage of lowly expressed genes where selection is absent. Over time this means lowly expressed genes might become progressively less efficiently translated since tRNAs are adapted to best translate the codon usage of highly expressed genes. At some critical point however, the inadequate codon usage of lowly expressed genes will impact greatly upon fitness, such that selection favours the duplication of alternative tRNA genes, which in turn generates the selective pressure for the identity of optimal codons to switch.

While it remains unclear which of these mechanisms has been most influential in shaping optimal codon divergence, comparative analyses of 160 bacterial species provides evidence

that codon usage, tRNA genes and base composition coevolve (HENRY 2007). A clear trend emerged for species' base composition to be reflected in third optimal codon position. For instance G+C ending codons were observed among the G+C rich Alpha Proteobacteria, whilst A+T ending optimal codons were observed among the A+T rich Firmicutes. Concordant optimal codons and base composition are consistent with either of the above mechanisms for optimal codon switching, with a role for mutational biases in shaping codon identities. Large scale switches in optimal codon identity were identified and mapped to the phylogeny (HENRY 2007). The 16 instances of optimal codon usage switching occurred in only five locations across the bacterial phylogeny, indicating coordinated switching events among amino acids. Furthermore, each of these switching events coincided with a change in tRNA gene content, consistent with the coevolution of codon usage and tRNA genes (BULMER 1987). A more recent study attempted to investigate variation in optimal codon usage (HERSHBERG and PETROV 2009) but (i) included species for which translational selection is absent and (ii) misidentified many optimal codons, and so the validity of their finding remain unclear.

1.1.2.3 The benefit of translational selection

Whilst it is clear that selection favours those codons which are best recognised by the most abundant tRNAs, the reason why optimal codons are advantageous remains a subject of debate. For many years it has been accepted that selection primarily acts to promote the efficiency of protein synthesis for rapid growth (ANDERSSON and KURLAND 1990; EHRENBERG and KURLAND 1984). By this mechanism, optimal codons are those translated more quickly than their synonyms, and so their usage increases the rate of translation elongation. Ribosomes spend less time translating each individual transcript and are thus more efficient – able to translate a larger numbers of mRNAs in a given time. An increase in translational efficiency is likely to increase growth rate (CONDON *et al.* 1995; EHRENBERG and KURLAND 1984).

More recently it has been suggested that selection for accurate translation is the major cause of selected codon usage bias (DRUMMOND and WILKE 2008; STOLETZKI and EYRE-WALKER 2007). Under the accuracy hypothesis, optimal codons are those translated with the lowest frequency of error. There are different ways in which the selective benefit of translational accuracy may be realised. It might be that misincorporated amino acid residues inflict a

metabolic cost (AKASHI 1994), and lead to the incorrect folding of protein products, potentially resulting in the formation of costly cytotoxic protein aggregates (WILKE and DRUMMOND 2006). Alternatively the benefit of translational accuracy might be an efficiency saving, either due to an increased yield of correctly translated protein products (BULMER 1991b), or due to a reduction in time spent proofreading (EHRENBERG and KURLAND 1984; LOVMAR and EHRENBERG 2006). Different classes of translation error are expected to have different effects upon correct protein function (EYRE-WALKER 1996). Missense errors may inflict the greatest cost to protein function when they occur at structurally and functionally important sites, and where misencorporated amino acids differ considerably in biochemical properties from the correct amino acids. Nonsense errors produce truncated and often non-functional proteins. Under an 'accuracy for efficiency' model, the cost of nonsense errors is likely to be greatest for codons downstream within gene sequences, since dysfunctional proteins are more costly in the case where more amino acids and translation time have been invested in their production.

The translation machinery is a very costly intracellular component, with ribosomes contributing two thirds of the protein mass of rapidly growing *E. coli* cells (PEDERSEN *et al.* 1978). It therefore seems likely that any processes improving the efficiency of translation will have large fitness consequences. It is mechanistically plausible that codons are selected for efficiency since it has been determined experimentally that some codons are translated more quickly than others (CURRAN and YARUS 1989; SØRENSEN and PEDERSEN 1991), including the C-ending codons of the two fold families of the form YNY which are always observed to be optimal (SHARP *et al.* 2005). Experimentally determined translation elongation rates appear to be faster for mRNAs which use optimal codons; *lacZ* constructs engineered in *E. coli* to contain non-optimal codons were translated around three seconds (4%) more slowly than those of the native form (SØRENSEN *et al.* 1989). There is also empirical work to suggest that selection is for translational efficiency. The interspecific correlation of the strength of selected codon usage bias with growth rate (SHARP *et al.* 2010) is most easily explained if selection is for efficiency, or 'efficiency mediated by accuracy' because the average time saved in translating optimal codons over non optimal codons is a greater proportion of the generation time of rapidly growing Bacteria, and so of greater selective benefit. One caveat to this argument is the assumption that the correlation of the strength of selected codon usage bias ($2N_es$) with growth rate reflects variation in the selection coefficient (s) among Bacteria. It might be that the correlation rather reflects variation in the effective population size (N_e), if

effective population size were correlated with growth rate, but this seems unlikely among unicellular organisms and as yet little is known about bacterial effective population sizes.

Translation error rates have been estimated to occur as frequently as 10^{-4} among wild-type *E. coli* cells (PARKER 1989) and so a substantial fraction of proteins are likely to contain at least one incorrectly incorporated amino acid. It has been demonstrated that some codons are more accurately translated than others (ORTEGO *et al.* 2007; PRECUP and PARKER 1987) meaning that codon usage has the potential to impact upon error rates. If the main selective benefit of accurate translation is in synthesising a higher fraction of functional or correctly folded proteins, it follows that there will be heterogeneity in the strength of selection for accurate codons among sites within genes, with selection of greatest intensity acting upon the most functionally and/or structurally important residues. Therefore associations of optimal codons and functionally conserved sites can be interpreted as evidence for translational accuracy (AKASHI 1994). By this rationale there appears to be some evidence of selection for accurate codons in *E. coli*. Whilst one study could not detect any association of optimal codons with conserved residues among *E. coli* and *S. enterica* (HARTL *et al.* 1994), it is likely that a large fraction of non-conserved residues were adaptive (CHARLESWORTH and EYRE-WALKER 2006) and so of similar functional importance as conserved residues. By defining conserved sites as those for which no amino acid polymorphisms were observed, an excess of optimal codons among conserved sites was detected (STOLETZKI and EYRE-WALKER 2007), although the magnitude of the effect was slight. Furthermore, accuracy-selected codon usage bias predicts a correlation between synonymous and nonsynonymous substitution rates; a correlation observed, and otherwise inadequately explained in *E. coli* (SHARP and LI 1987b) and *Buchnera aphidicola* (TOFT and FARES 2009). Weak associations of optimal codon usage with both structurally important residues (ZHOU *et al.* 2009) and sporadic processing by the molecular chaperone GroEL (WARNECKE and HURST 2010) provide further evidence that accuracy has some role in translational selection, although it remains unclear whether translational accuracy is the dominant target of codon selection.

It is possible that selection is for both the accuracy and efficiency of translation provided that efficient codons are also accurate. In this case codon selection would allow species a rare evolutionary opportunity to overcome the observed trade-off of growth rate and growth yield (MIKKOLA and KURLAND 1992), with seemingly no added cost. Early experimental work however suggested that efficient codons were not accurate. For the amino acid

phenylalanine, it was observed that the optimal UUC codon was most efficient, whilst the non-optimal UUU codon was most accurate (DIX and THOMPSON 1989; PARKER and PRECUP 1986) however these studies were only able to consider a restricted class of error substrates, and appear to be inconsistent with the (slight) excess of UUC synonyms observed among conserved sites (STOLETZKI and EYRE-WALKER 2007). In conclusion, selection most likely proceeds via the efficiency and accuracy of translation but for now it remains unclear which of these is more important in Bacteria.

1.1.3 Other forms of selected codon bias

Whilst selection for translationally optimal codons is the major source of selected codon bias within and among bacterial genomes, there is evidence that selection of other forms impacts upon the evolution of synonymous sites. Enrichment of non-optimal A+T rich codons has been observed at the start (first ~30bp) of genes in *E. coli* (EYRE-WALKER and BULMER 1993) and *H. pylori* (LAFAY *et al.* 2000). In *E. coli* these sites exhibit reduced synonymous divergence consistent with the action of purifying selection for an alternative form of selected codon bias. Recent work indicates that these sequences are a common feature of genes; consistently exhibiting A+T richness and low mRNA stability in a wide range of organisms (GU *et al.* 2010; TULLER *et al.* 2010; TULLER *et al.* 2009). Experimental work has demonstrated that the mRNA stability of these regions is the most important determinant of gene expression level among artificially engineered GFP constructs (KUDLA *et al.* 2009) and implies that these sequences may have a role in translation initiation. Site specific codon preferences have been observed among genes in *E. coli* meaning the identity of optimal codons may depend on the identities of neighbouring codons (BULMER 1990; MAYNARD SMITH and SMITH 1996). For example the codons GUU and GUA are both optimal for the amino acid valine in *E. coli* (SHARP *et al.* 2010) and yet GUA is used with a much lower frequency when followed by the nucleotide bases A or G. It has been suggested that some of these effects are explained by the avoidance of AGG sequences and out-of-frame stop codons (MAYNARD SMITH and SMITH 1996). Some context effects might reflect more complex patterns of translational selection; the occurrence of codon pair preferences has been well documented (HATFIELD *et al.* 1992; IRWIN *et al.* 1995) and may bear some biological relevance given *in vivo* rate variation in translation elongation among codon pairs (WEN *et al.* 2008). Various other patterns of codon usage bias have been observed but are of small effect and are not discussed.

1.2 Codon usage bias in Eukaryotes

Here I review what is known of codon usage in Eukaryotes, with particular emphasis upon unicellular organisms.

1.2.1 Codon usage bias in unicellular Eukaryotes

The same forces of natural selection and mutation shape patterns of codon usage among unicellular Eukaryotes as have been exemplified for Bacteria. Many unicellular Eukaryotes vary in their G+C content; known examples vary from 18% in *Plasmodium falciparum* (GOMAN *et al.* 1982) to 63% in *Giardia lamblia* (LAFAY and SHARP 1999), and so presumably variation in codon usage among species can be attributed to alternative patterns of mutation. Variation in codon usage among genes within genomes has been attributed to G+C content in many species, including the apicomplexan parasites *P. falciparum* (PEIXOTO *et al.* 2004) and *Cryptosporidium parvum* (GROCOCK and SHARP 2001). In *C. parvum*, the trend may reflect a regional mutational bias since GC3s among genes is correlated with G+C content of respective 3′ and 5′ flanking sequences. Regional variation in GC3s has also been reported for the yeast *Saccharomyces cerevisiae* (SHARP and LLOYD 1993). So, unlike the situation observed for Bacteria, regional mutational biases in G+C content may be a common feature of unicellular Eukaryotes.

Natural selection appears to influence patterns of codon usage in an extremely diverse range of species including the Ascomycota *Saccharomyces* (BENNETZEN and HALL 1982; SHARP *et al.* 1988), *Aspergillus nidulans* (LLOYD and SHARP 1991) and *Kluyveromyces lactis* (LLOYD and SHARP 1993), the mycetozoan *Dictyostelium discoideum* (SHARP and DEVINE 1989), the Euglenozoa *Trypanosoma*, *Leishmania*, and *Crithidia* (ALVAREZ *et al.* 1994; HORN 2008), the diplomonad *Giardia lamblia* (LAFAY and SHARP 1999), the amoebozoan *Entamoeba histolytica* (ROMERO *et al.* 2000), the Ciliophora *Tetrahymena thermophila* and *Paramecium tetraurelia* (SALIM *et al.* 2008), the Apicomplexa *P. falciparum* and *C. parvum* (GROCOCK and SHARP 2001; PEIXOTO *et al.* 2004), and the chlorophytan *Chlamydomonas reinhardtii* (NAYA *et al.* 2001). The first evidence for natural selection in Eukaryotes came from the yeast *S. cerevisiae* where certain codons were overrepresented in highly expressed genes (BENNETZEN and HALL 1982) and were complementary to the most abundant tRNAs (IKEMURA 1982). For several Eukaryote species with A+T rich genomes (*D. discoideum*, *T. thermophila*, *E. histolytica* and *C. parvum*), it has been noted that the identities of optimal codons are almost exclusively G+C

ending, and so selection opposes the prevailing direction of mutational pressure. These observations contrast with patterns of optimal codon usage observed among Bacteria, where optimal codon identity is typically shaped by mutational biases (HENRY 2007) but not without exception (e.g. MUSTO *et al*. 2003), and so it remains unclear whether this is a genuine difference between Bacteria and Eukaryotes.

The strength of selected codon bias has been estimated for a number of Eukaryotes (DOS REIS and WERNISCH 2008) using the method of Sharp and collegues (2005; see 1.1.2.2). For the yeast species *S. cerevisiae*, *Cryptococcus neoformans* and *Neurospora crassa* selection appears to be very strong ($S > 2$), whilst in *P. falciparum* and the microsporidium *Encephalitozoon cuniculi*, S values are significant but much lower ($S = 0.66$ and 0.56 respectively). These values, expected to correspond to $2N_eS$ in haploids and $4N_eS$ in diploids, are similar in range and magnitude to estimates obtained for Bacteria, and the reason for this is not quite clear. Effective population sizes have been estimated to be approximately an order of magnitude lower in unicellular Eukaryotes where $N_e \sim 10^7$ (LYNCH *et al.* 2006), than in Bacteria where $N_e \sim 10^8$ (LYNCH 2007), and perhaps implies that the selection coefficient upon codon bias may be much higher for unicellular Eukaryotes, and there is no obvious reason for this. As such, a distinction should be made between the asexual lives of Bacteria and those of unicellular Eukaryotes which are punctuated by sexual recombination. Whilst some recombination is evident in many Bacteria with the occurrence of horizontally transferred genes, it is less clear, the frequency at which *homologous* recombination occurs. By reducing linkage among sites, homologous recombination can improve the efficacy of selection analogous to increasing the effective population size (HILL and ROBERTSON 1966). Since homologous recombination is implicit to meiosis, it is likely to occur at a higher frequency among unicellular Eukaryotes than among Bacteria, and might go towards explaining the similar values of $2N_eS$ (KAISER and CHARLESWORTH 2009; SHARP *et al.* 2010).

The occurrence of diploid phases in the lifecycles of unicellular Eukaryotes allows for the occurrence of a third evolutionary process, biased gene conversion; for review see MARAIS (2003). Several mechanisms of DNA repair are known to proceed via gene conversion events, whereby a damaged DNA strand is invaded by its undamaged homologue and used as a template for repair. There is substantial evidence that DNA repair mechanisms are biased in favour of converting mismatched bases (e.g. T:G) into G and C bases (BROWN and JIRICNY 1989). So following the initiation of gene conversion repair, any heterozygous alleles

within the invaded strand will have a tendency to be "repaired" to the G:C homozygous state. The effect of biased gene conversion is to therefore distort the segregation of alleles, thus acting like a selection coefficient in favour of G and C alleles (NAGYLAKI 1983). There is considerable evidence that biased gene conversion influences patterns of codon usage bias in *S. cerevisiae*. First it was observed that GC3s is correlated with recombination rate among genes (BIRDSELL 2002); an observation expected if biased gene conversion were at work since conversion events are induced by recombination. Conclusive evidence for biased gene conversion in *S. cerevisiae* came from the examination of the meiotic products of a highly polymorphic hybrid (MANCERA *et al.* 2008). It was revealed that ~1% of the yeast genome is subject to a gene conversion event at each meiosis, and that gene conversion products are 1.4% more G+C rich than the SNP base content in the parental genomes. To determine the impact of biased gene conversion upon synonymous codon usage bias, a recent study has estimated the strength of biased gene conversion (S_{BGC}) by comparing codon frequencies of low expression low recombination genes (assumed to evolve neutrally) with low expression high recombination genes (assumed to be influenced by biased gene conversion) (HARRISON and CHARLESWORTH 2010). This indicated that biased gene conversion has a relatively minor role shaping the codon usage of yeast (*S. cerevisiae* $S_{BGC} = 0.09$) compared with selection upon codon usage bias (*S. cerevisiae* $S = 1.12$).

1.2.2 Codon usage bias in multicellular Eukaryotes

There is no clear distinction between the codon usage of unicellular and that of multicellular Eukaryotes. Multivariate analyses of the codon usage across five model Eukaryotes reveals a single major trend associated with variation in genome GC3s (KANAYA *et al.* 2001) consistent with variation in mutational biases. There is also evidence that natural selection can shape patterns of codon usage within the genomes of certain multicellular Eukaryotes. In *Caenorhabditis*, codon usage is distinctive among highly expressed genes (STENICO *et al.* 1994), corresponding to both predicted (DURET 2000) and experimentally determined (KANAYA *et al.* 2001) tRNA abundances. The same situation is observed for *Drosophila* (AKASHI *et al.* 1998; MORIYAMA and POWELL 1997; SHIELDS *et al.* 1988). Estimates of the strength of selection for these species are ~1 and thus similar in magnitude to estimates for unicellular Eukaryotes and Bacteria (CUTTER 2008; DOS REIS and WERNISCH 2008; MASIDE *et al.* 2004) despite much lower effective population sizes where $N_e \sim 10^6$ (CUTTER *et al.* 2006; NEI and GRAUR 1984) and so widen the apparent paradox described above. There is evidence that biased gene

conversion is at work in *Drosophila* since an excess of GC → AT polymorphisms are observed in non-coding regions (HADDRILL and CHARLESWORTH 2008). The strength of this putative biased gene conversion has been estimated as $S_{BGC} = 0.6$ in regions of high GC content (HADDRILL and CHARLESWORTH 2008), and suggests that biased gene conversion has a more substantial impact upon patterns of codon bias in Drosophila than in yeast. The observation is consistent with low rates of sexual reproduction and high rates of inbreeding in yeast, which reduce the effectiveness of biased gene conversion (HARRISON and CHARLESWORTH 2010).

A distinction can be made with patterns of codon usage in mammalian genomes, where there is substantial variation in GC3s among genes (BERNARDI *et al.* 1985; SHIN-ICHI and IKEMURA 1986). In humans GC3s varies from 30% to 90% among genes (SHIN-ICHI and IKEMURA 1986) and regional patterns in variation (termed isochores) occur over a much larger scale (~300kb) than is observed among non-mammalian genomes (FILIPSKI *et al.* 1973). The cause(s) of isochores remain unclear; for review see Eyre-Walker and Hurst (2001), although there seems to be no role for translational selection (SÉMON *et al.* 2005). There is growing evidence that the evolution of isochores could be driven by biased gene conversion (DURET and GALTIER 2009; EYRE-WALKER 1999; GALTIER and DURET 2007) which might explain the relatively rapid timescales over which these structures evolve (DURET and ARNDT 2008). So whilst it is clear that mutational biases and probably biased gene conversion shape patterns of codon usage within mammalian genomes, it is less clear that there is any role for natural selection. Nevertheless some authors have estimated the strength of selection for humans to obtain values which are low but greater than zero (DOS REIS and WERNISCH 2008; SUBRAMANIAN 2008); however it is possible that these values reflect other forms of selection of small effect (see CHAMARY *et al*. (2006)) or the failure to correct for heterogeneity in base composition.

1.3 Codon usage bias in Archaea

From the first section it is clear that much is known about synonymous codon usage in Bacteria and some groups of Eukaryotes. The Archaea are a monophyletic and diverse group of organisms, estimated to have shared common ancestry with Bacteria more than four billion years ago (BATTISTUZZI *et al.* 2004). In stark contrast with Bacteria, little is known about the codon usage of Archaea since early genome sequencing primarily focused upon

medically and industrially relevant Bacteria. Archaea inhabit a diverse range of environments including extremes of temperature, pressure and salinity. The quest to understand ecology and adaptation in these environments and their commercial applications means that there are now genome sequences available for nearly 70 species of Archaea. Yet despite this plethora of sequence data, few studies have examined synonymous codon bias in Archaea.

Several investigations of individual genome sequences have incidentally examined patterns of codon usage in Archaea. Firstly, in a study of three of the first available microorganism genome sequences, correspondence analysis of relative synonymous codon usage (RSCU) values was performed in an attempt to investigate trends in codon usage bias among the genes of *Methanococcus jannaschii* (MCINERNEY 1997). Such analyses of RSCU values have been demonstrated to be prone to erroneous results (PERRIERE and THIOULOUSE 2002), and the study of *M. jannaschii* appears to be one such victim since it inadvertently detected a primary trend associated with the usage of the amino acid proline. Thus, sources of variation in synonymous codon usage among genes in *M. jannaschii* remain unclear. Another study examined patterns of synonymous codon usage bias within the archaeon *Nanoarchaeum equitans* (DAS *et al.* 2006). A rather low level of heterogeneity in synonymous codon usage was observed among genes; nevertheless a primary trend explaining this variation was identified and was found to be correlated with a measure of selected codon usage bias (CAI; $r = 0.52$), consistent with the action of translational selection. Finally, comparative analyses of the archaeal species *M. jannaschii* and *Methanococcus maripaludis* attempted to identify patterns of divergence and codon usage bias among different functional categories of genes. The synonymous substitution rates reported revealed that more than three substitutions per site had occurred since the common ancestor of *M.jannaschii* and *M. Maripaludis,* meaning that synonymous sites are saturated with changes and cannot not be reliably compared. Correspondence analysis of RSCU values was performed to identify trends in codon usage bias among the genes of both species, but as with the McInerney report, the resulting trends seem to be obscured by the effects of biased amino acid composition, and so it seems that no conclusions can be drawn from the study.

There are no large scale studies of synonymous codon usage bias in Archaea, however several studies primarily examining variability in codon bias among Bacteria have included Archaea in their analyses. For instance the Lynn *et al*. (2002) and Chen *et al*. (2004) studies

(discussed) contained 8/40 and 10/100 archaeal species respectively. Similarly, a study which tested for the presence of translational selection within genomes based upon inferred coadaptation of codon usage and tRNA gene content, included 12 species of Archaea (DOS REIS *et al.* 2004). Of these, six were found to exhibit significant levels of selected codon usage bias, although this measure of selected codon usage bias did not correlate well with a population genetics-based approach (SHARP *et al.* 2005). A recent examination of the impact of various genomic characteristics upon growth rate (VIEIRA-SILVA and ROCHA 2010) included 26/214 species of Archaea and concluded that there were no differences in the evolutionary processes governing bacterial and archaeal genomes. Their reasoning was that they observed no significant difference in the deviations of Archaea and Bacteria from a linear model describing growth rate as a function of various genome characteristics. Yet the distribution of archaeal datapoints does not appear to be randomly distributed (see Figure 1 in VIERA-SILVA AND ROCHA 2010), and so their conclusion may be an oversimplification. The study suggests that translational selection is present in some archaeal genomes but the significance of this was not tested. Thus little is known of the role of natural selection in shaping patterns of codon usage bias among Archaea, and so it is surprising that one study has attempted to use patterns of codon usage bias to predict gene expression level (KARLIN *et al.* 2005) when differential codon usage among highly expressed genes primarily reflects translational selection. The failure to first test for the presence of translation selection prior to the prediction of highly expressed genes means that predictions for any species where selection is absent are likely to be invalid. Several other problems have been identified with their approach (HENRY and SHARP 2007) and so it does not seem possible to draw conclusions from their report. In conclusion, there is some evidence for the presence of translational selection in the genomes of some Archaea but our overall knowledge of patterns of codon bias in Archaea is somewhat limited.

1.4 Aims

Since much is known about codon usage in Bacteria and many Eukaryotes, and yet little is known about codon usage in Archaea, this study seeks to rectify the taxonomically biased distribution of knowledge. There are reasons to suspect *a priori* the codon usage of Archaea may be rather different from that of Bacteria. First, the numbers of ribosomal RNA operons and tRNA genes, known to be correlated with the strength of selected codon usage bias, are much lower in range among Archaea than Bacteria. So if Archaea follow similar trends to

Bacteria, then they are expected to exhibit a reduced intensity of selected codon bias. Second, unlike Bacteria, Archaea primarily inhabit extreme environments. Since numerous reports have claimed to identify ecological associations with synonymous codon usage, analysis of codon usage in Archaea will provide a wider range of datapoints to confirm or refute these reports, and to determine exactly how ecology in extreme environments impacts upon codon bias.

Each of the chapters seeks to address a different question. The third chapter aims to deduce the major trends in synonymous codon usage in a well-studied archaeon, *Methanococcus maripaludis*, and compare them with patterns observed in the model bacterium *Escherichia coli*. The fourth chapter extends these analyses, identifying the factors underlying major trends in codon usage bias within the genomes of a further 66 species of Archaea. The fifth chapter aims to estimate the strength of selected codon usage bias for each species of Archaea and explores how it varies by comparison with bacterial genomes. The sixth chapter seeks to discern major trends explaining variation in codon usage bias among the genomes of Archaea. The seventh chapter determines the identities of optimal codons for each species of Archaea, exploring if and how their identities and patterns of selection vary across different amino acid groups. Finally, the eighth chapter attempts to partition the total strength of selected codon usage bias estimated for each species of Archaea into separate components due to the accuracy and efficiency of translation.

2. METHODS

Described here are methods common to many of the results chapters; here they are provided primarily for reference.

2.1 Indices of codon usage bias

The indices described here are useful in summarising aspects of codon usage. Simple indices may be computed for any coding sequences (1-5). Other indices describing patterns of selected codon usage bias (6-7) require more information since the identities of some optimal codons vary among species.

2.1.1 G+C content at synonymously variable third codon positions (GC3s)

GC3s is the proportion of all synonymous codons (i.e. all codons except Trp, Met and stop codons) that are G and C ending:

$$\mathrm{GC3s} = \frac{[(\mathrm{NNS}) - (\mathrm{AUG} + \mathrm{UGG} + \mathrm{UAG})]}{[\mathrm{NNN} - (\mathrm{AUG} + \mathrm{UGG} + \mathrm{UAA} + \mathrm{UAG} + \mathrm{UGA})]} \qquad (1)$$

Where N = any nucleotide base and NNS = NNC + NNG.

Values of GC3s among genes are expected to follow a binomial distribution, meaning that the standard deviation of GC3s among genes is expected to be greatest at 50% G+C content. The expected standard deviation of GC3s can be computed as:

$$SD\ \mathrm{GC3}s = \sqrt{\frac{p\ (1-p)}{l}} \qquad (2)$$

Where p is the mean GC3s and l is the harmonic mean of the number of synonymously variable third codon positions among genes.

2.1.2 Additional base composition indices of synonymously variable third codon positions (X3s)

Other indices of synonymous base composition are useful for examining strand specific mutational patterns and take the same form:

$$\mathrm{X3s} = \frac{\mathrm{NNXs}}{\mathrm{NNNs}} \qquad (3)$$

Where X = base of compositional interest e.g. G or G+T, N = any nucleotide base, and s indicates that the fraction is restricted to synonymous codons.

2.1.3 Base composition skew indices

Skew indices typically measure the enrichment of G over C content, T over A content, or combined G+T over C+A content at synonymously variable third codon positions across genes, and may be used to examine strand specific mutational biases (LOBRY 1996; MCLEAN *et al.* 1998). The cumulative sum of skew values across the genes of a chromosome is often used to predict the location of replication origins in bacterial genomes; however these analyses are often ambiguous when applied to archaeal genomes due to the occurrence of multiple origins of replication per chromosome (e.g. LUNDGREN *et al*. 2004). Skew values in these analyses are computed in the form:

$$\text{X-Y skew} = \frac{(\mathrm{X3s} - \mathrm{Y3s})}{(\mathrm{X3s} + \mathrm{Y3s})} \qquad (4)$$

Where X and Y are the bases of compositional interest e.g. X = G and Y = C across synonymously variable third codon (3s) positions.

2.1.4 Relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) values may be used to compare the codon usage of alternative degeneracy families, removing biases in amino acid composition (SHARP and LI 1986). RSCU values are computed as the codon usage observed divided by that expected under uniform codon usage (i.e. divided by the average codon usage for that amino acid group). An RSCU value of >1 indicates that a codon is used more frequently than expected under random usage; conversely a value of <1 indicates usage occurring less frequently than random. Deviations from uniform codon usage may be due to mutational biases or natural selection.

$$\mathrm{RSCU}_{ij} = \frac{\mathrm{X}_{ij}}{\frac{1}{n_i}\sum_{j=l}^{n_i}\mathrm{X}_{ij}} \qquad (5)$$

Where X_{ij} is the j^{th} codon usage for the i^{th} amino acid from a total of l synonymous codons and n_i is the degeneracy of the i^{th} amino acid

2.1.5 The effective number of codons (N_c)

The effective number of codons (N_c) is a measure of general codon usage bias, with values ranging from 61 under uniform codon usage, to 20 in the case of extreme codon bias, whereby only a single codon is used for each amino acid (WRIGHT 1990). The index draws an analogy between the codons within a gene and the alleles within a population. Then, existing population genetic indices of homozygosity may be applied to the codon usage of a gene to indicate its degree of codon bias. First, codon usage homozygosity (F) is estimated for each i^{th} amino acid group:

$$F_i = \frac{\left(n_i \sum_{j=l}^{k} p_j^2 - 1\right)}{(n_i - 1)} \qquad (6)$$

Where p_j is the frequency of the j^{th} codon from a total of n_i sites for that k-fold degenerate i^{th} amino acid across l synonymous sites. The mean homozygosity (F) for each k-fold i^{th} amino acid is F_k.

Then, N_c is a function of the mean homozygosities across all k degeneracy families:

$$\mathrm{N_c} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \qquad (7)$$

Low N_c values indicate restricted codon usage bias which can reflect mutational or selective forces. Wright (1990) developed a graphical method to examine the distribution of N_c values with respect to G+C content related mutational biases using plots of N_c against GC3s (e.g. Figure 3-1). Expected N_c values are expected to vary with GC3s as follows:

$$N_c = 2 + S + \left[\frac{29}{S^2 + (1 - S)^2} \right] \quad (8)$$

Where S = the G+C content at synonymously variable third codon positions (GC3s).

Plots of N_c against GC3s can be inspected to identify distributions of genes with lower N_c values than expected, and in conjunction with other analyses, may be interpreted as reflecting natural selection (e.g. Figure 3-1).

Some authors have used observed-to-expected scaled N_c values (NOVEMBRE 2002) denoted N_c' or ENC' as indices of selected codon usage bias (e.g. HERSHBERG and PETROV 2009; ROCHA 2004; SUBRAMANIAN 2008), however there might be problems with this approach since N_c is not specific to translational selection and so also encapsulates bias from mutational processes. The index typically uses the current synonymous base composition of a gene as an indication of that expected in the absence of selection. Since selected codon usage bias can impact upon base composition, the method is expected to systematically underestimate selection where optimal codons are G+C ending and overestimate selection where optimal codons are A+T ending. Additionally, implicit to the method is an assumption of gene-specific mutational biases, for which there is no evidence among bacterial or archaeal genomes. Given these potential issues, scaled N_c values were not used in these analyses.

2.1.6 The frequency of optimal codons (F_{op})

The frequency of optimal codons is a measure of selected codon usage bias which can be used to compare genes within species provided that there is no large variation in mutational bias among genes. Since the identity of translationally optimal codons varies among species, it is first necessary to determine their identity (see 2.2) before proceeding to calculate their frequency. Then, F_{op} is simply the proportion of all codons that are optimal.

$$\mathrm{F_{op}} = \frac{n_{opt}}{n_{opt} + n_{non-opt}} \tag{9}$$

Where n_{opt} is the number of optimal codons and $n_{non\text{-}opt}$ is the number of non-optimal codons.

2.1.7 The codon adaptation index (CAI)

The codon adaptation index (CAI) is another species-specific measure of selected codon usage bias, differing from F_{op} in assigning different weight to different suboptimal codons. Codon weight is assigned as fitness (w) values, corresponding to the codon usage relative to the most frequently used codon for an amino acid group across a set of highly expressed genes:

$$w_{ij} = \frac{n_{ij}}{n_{i\,max}} = \frac{\mathrm{RSCU}_{ij}}{\mathrm{RSCU}_{i\,max}} \tag{10}$$

Where n_{ij} is the number of occurances of the j^{th} codon of the i^{th} amino acid across a set of highly expressed genes and $n_{i\,max}$ is the maximum number of codons of any given synonymn for that i^{th} amino acid.

Then, CAI values can be computed as the geometric mean of fitness values for the codons of a gene:

$$\mathrm{CAI} = \frac{1}{l}\sum_{j=1}^{l} \ln w_j \tag{11}$$

Where w_j is the fitness value of the j^{th} codon among a total of l synonymous codons within a gene

In assigning fitness values the CAI assumes that sequences of highly expressed genes are largely influenced by natural selection, and so is meaningless for species where translational selection is absent. There are also issues for the CAI in species where codon usage is dominated by strong mutational biases (GROCOCK and SHARP 2002). For example, the codon usage in *Pseudomonas aeruginosa* is influenced by strong mutational pressure in the direction of A+T → G+C, and yet it is the T-ending codon that is optimal for the amino acids glycine and alanine. For these amino acids, while the T-ending synonymns are overrepresented across highly expessed genes relative to other genes (and thus optimal), it is in fact the C-ending synonynms that are used with the highest frequency across highly expressed genes due to a strong mutational pressure. Since CAI fitness values are assigned based only upon the codon usage across highly expressed genes (equation 10), then for these amino acids, the non-optimal C-ending codons are assigned CAI fitness values of 1, and the T-ending codons are both assigned fitness values of 0.62. Thus, a gene containing only optimal codons for these amino acids would have a F_{op} value of 1 but a CAI of only 0.62. Similarly, a gene comprised only of non optimal C-ending codons for these amino acids would possess a F_{op} value of 0 and yet a CAI value of 1.

2.2 Within-group correspondence analysis

Multivariate analyses are commonly used to explore variation in codon usage among genes (KANAYA *et al.* 2001; KLOSTER and TANG 2008; LOBRY and CHESSEL 2003; LOBRY and NECSULEA 2006; LYNN *et al.* 2002; MCINERNEY 1997; SUZUKI *et al.* 2008). Within-group correspondence analysis (WCA) allows synonymous codon usage to be explored independently of amino acid compositional biases (LOBRY and CHESSEL 2003). From the codon usage of the 59 synonymous codons, WCA identifies 41 orthogonal axes (corresponding to the remaining degrees of freedom with 18 amino acid groups) which successively explain the most variation among genes. The biological relevance of each axis can be assessed by considering the axis positions of known genes and/or codons as well as the correlation of these positions with simple codon usage statistics. Only the first three axes are typically of biological relevance for any bacterial genome, and any given genome may show some or all of the sources of variation seen across all Bacteria (SUZUKI *et al.* 2008).

2.3 Identifying optimal codons

Optimal codons were identified as those significantly over represented in a dataset of highly expressed genes relative to a dataset of genome-wide codon usage by Pearson chi-squared tests with sequential Bonferroni correction, as assessed by Henry and Sharp (2007).

2.4 Estimating the strength of selected codon usage bias (*S*)

The strength of selected codon usage bias (*S*) is a measure of codon bias due to natural selection which controls for codon bias resulting from unequal mutational patterns, allowing for interspecific comparisons of translational selection (SHARP *et al.* 2005). The index is based upon a population genetic model of selection-mutation-drift equilibrium (BULMER 1991b), which considers the case of two synonyms, one of which is optimal and thus selectively favoured over the other, and where there may be an unequal mutation rate from one codon to the other. This is the same situation as that first described by Kimura (1983, p. 43), who considered the more general case of two alleles within a population. By assuming that species are always near fixation and that the ratio of transition rates is an exponential function of $2N_eS$ (where N_e is the effective population size and s is the selection coefficient), it is possible to arrive at an approximation of the strength of selection (S), corresponding to the haploid population genetic parameter $2N_eS$, as a function of optimal codon equilibrium frequencies under (i) the combined effects of natural selection, mutational biases and random genetic drift (P_s) and (ii) neutral processes alone (P_n):

$$S = ln\left[\frac{P_s\,(1-P_n)}{P_n(1-P_s)}\right] \qquad (12)$$

The strength of selection (S) may then be estimated by using optimal codon frequencies from a dataset of highly expressed genes as an estimate of P_s, and optimal codon frequencies across the entire genome as an estimate of P_n. For the majority of these analyses, and unless stated otherwise, optimal codon frequencies were restricted to the C-ending optimal codons for the amino acids Phe, Tyr, Ile and Asn, which with the exception of Ile, are two-fold degenerate amino acid groups. All four of these amino acid groups are decoded by a single tRNA species and so best conform to the original model (BULMER 1991b), with apparently invariant optimal codon identities among species. The threefold degenerate amino acid Ile was treated as a twofold degenerate amino acid with the exclusion of the rarely used AUA codon. Values of S expected by chance (centered on zero) were obtained from the

distribution of S values from 1000 replicates of randomly assigned highly expressed genes (when sampling without replacement).

In further analyses, S was estimated for the codons of four-fold and six-fold degeneracy families. In these cases, the equilibrium frequencies (P_s and P_n) are estimated from the frequency of the optimal codon relative the total number of all codons for that amino acid. This approach of pooling multiple non-optimal codons into a single non-optimal codon allelic class is only accurate in the case where all non-optimal codons have the same selection coefficient, although the degree of error from this approach does not appear to be large.

Note that the strength of selected codon usage bias (S) is an estimate of the compound parameter $2N_eS$ (in the case of haploids), and not the selection coefficient upon codon bias (s). It only approximates estimates of $2N_eS$ in the case where the codon frequencies of highly expressed genes are at selection-mutation-drift equilibrium, and where those of lowly expressed genes are at mutation-drift equilibrium. The approach assumes that the selection coefficient upon codon usage (s) does not vary in magnitude or direction among sites. Any such heterogeneity is likely to result in an underestimation of the point estimates of S across the majority of sites because the presence of a minority of very weakly selected sites, neutral sites and or sites with selection in the opposite direction, will tend to increase the frequency of 'non-optimal codons'. Under the model, these non-optimal codons are taken to reflect the inability of selection to fully spread the optimal codons to fixation, but this would not be the case if there were site-specific codon preferences such that the 'non-optimal codon' is selectively favoured. Neverthless S indicates the extent of bias towards translationally optimal codons across high vs low expression genes.

2.5 Accounting for phylogenetic non-independence

The method of phylogenetic independent contrasts (FELSENSTEIN 1985) was used to assess correlations of various traits across Archaea whilst accounting for their underlying patterns of relatedness (estimated in Chapter 5). Contrasts for a given trait may be estimated by scaling the difference in trait values among pairs of species at the tips of the tree by the square root of the phylogenetic distance (i.e. expected variance) between them:

$$C_{ij} = \frac{x_i - x_j}{\sqrt{d_{ij}}} \qquad (13)$$

Where C is the contrast across the pair of taxa i and j, with trait values x_i and x_j and separated by a total branch length of d.

Further contrasts can be obtained by iterating the procedure across the deeper branches of the tree, using ancestral states inferred as the mean trait value of species pairs. As such, contrasts are obtained across all independent internal branches of the tree, so that the total numbers of contrasts obtained is equal to the numbers of internal nodes in the tree. The procedure of scaling by the expected variance can be inaccurate when traits evolve more quickly or more slowly than expected by a random walk, and may lead to some contrasts with disproportionate weight (GARLAND *et al.* 1992). To avoid this problem, branch lengths were transformed to obtain contrasts whereby there was no relationship with the absolute value of the contrast and its variance. Finally, regression analyses, forced through the origin (GARLAND *et al.* 1992), were used to assess trends among contrasts.

3. THE STRENGTH AND PATTERN OF SELECTED CODON USAGE BIAS IN *METHANOCOCCUS MARIPALUDIS*

3.1 Introduction

Here, patterns of synonymous codon usage bias are investigated in one of the most extensively studied archaeal species, the anaerobic methanogen *Methanococcus maripaludis*. The availability of genome-wide sequence (HENDRICKSON *et al.* 2004) and expression data (XIA *et al.* 2006) makes it a particularly convenient organism in which to examine patterns of codon usage bias. *M. maripaludis* is representative of the Archaea with a typical doubling time of 2.3 hours at 37°C (JONES *et al.* 1983), 3 rRNA operons and a minimal set of 37 tRNA genes. Among bacterial genomes, such characteristics are associated with weak or absent translational selection. Yet here we observe very strong selected codon usage bias in *M. maripaludis*, although patterns of selected bias differ from those in the bacterium *Escherichia coli*. In *M. maripaludis* selection is much weaker among four-fold degenerate sites.

3.2 Methods

3.2.1 Data sources

Protein coding sequences for the genome of *Methanococcus maripaludis* strain S2 (HENDRICKSON *et al.* 2004) were obtained from the GenBank database (accession number BX950229) using the ACNUC retrieval system (GOUY *et al.* 1985). Six genes shorter than 50 amino acids in length were excluded from subsequent analyses. There were 1716 protein coding genes in the final genome-wide dataset. A set of translation machinery-encoding genes expected to be highly expressed was identified on the basis of genome annotation and included ribosomal protein and elongation factor genes. To be confident that these genes were functional and highly expressed, we implemented two criteria: (1) that each gene must be present as an orthologue in at least 95% of 53 archaeal genome sequences, and (2) that the protein occurred among the top 25% of expression levels in a recent study (XIA *et al.* 2006). Six genes in total were excluded; three on the basis of absence in many genomes (the ribosomal protein genes LX, L30e and L34e), and three due to low protein abundance (L24a, EF-Tu domain 2 & EF-1b). There were 51 ribosomal protein genes in the final high expression dataset. Expression level data were available for 967 of the 1716 protein coding

Note: the manuscript of this chapter has been accepted for publication (see Appendices)

genes and were taken as the signal intensity (n1 values) of protein abundance data (XIA *et al.* 2006) normalised by protein molecular weight. The numbers of tRNA genes and predicted anticodons were obtained from the tRNA scan SE database (LOWE and EDDY 1997).

3.2.2 Exploring variation among genes

Variation in patterns of codon bias among genes was explored using two techniques commonly used to investigate variation in codon usage bias among genes: (i) a plot of the effective number of codons, N_c (WRIGHT 1990) against the G+C content at synonymously variable third codon positions, GC3s and (ii) within block correspondence analysis (WCA) (LOBRY and CHESSEL 2003). Values for N_c and GC3s were computed for each of the 1716 genes using codonW (Peden, J.F.). Within block correspondence analysis was implemented in the ade4 package (CHARIF *et al.* 2005) of the R suite.

3.2.3 Indices of selected codon usage bias

Relative synonymous codon usage values (RSCU) allowed for the convenient examination of the codon usage for highly expressed genes and genome wide datasets (SHARP and LI 1986). Optimal codons were identified as those occurring significantly more frequently in the highly expressed gene set. Two species-specific measures of selected codon usage bias, the frequency of optimal codons, F_{op} (IKEMURA 1981), and the codon adaptation index, CAI (SHARP and LI 1987a) , were computed in codonW to explore variation in selected codon usage bias among genes. Codon fitness values to determine CAI were estimated from codon usage in the 51 highly expressed genes dataset.

The strength of selected codon usage bias (S) was estimated for the highly expressed genes of *M. maripaludis* and *E. coli*. Confidence intervals were determined by permutation (Chapter 2). The highly expressed gene set for *E. coli* was comprised of 37 ribosomal protein genes and three elongation factor genes previously shown to exhibit selected patterns of codon usage bias (SHARP *et al.* 2005) and was not entirely homologues to the *M. maripaludis* dataset due to divergence of ribosomal protein genes (LECOMPTE *et al.* 2002). Nevertheless both highly expressed gene datasets possess analogous functions and so are likely to be influenced by a similar intensity of translational selection. To obtain estimates of S for four-fold degenerate amino acids, non-optimal codons were grouped into a single allelic class so that frequencies of optimal and non-optimal codons were compared. To ensure these values could be reliably

compared within and among species, amino acid groups for which selection was absent in either species (Cys and Glu, Lys), or for which tRNAs could not be directly compared (Gln) were excluded. The strength of selection was compared between degeneracy groups within species using t-tests and between species within degeneracy groups using paired t-tests.

		High		All				High		All				High		All				High		All	
Phe	UUU	72	0.75	16089	1.56	**Ser**	UCU	28	0.45	6178	1.21	**Tyr**	UAU	20	0.21	9940	1.06	**Cys**	UGU	33	0.85	4047	1.21
Phe	UUC*	121	1.25	4522	0.44	**Ser**	UCC	42	0.67	2430	0.48	**Tyr**	UAC*	171	1.79	8795	0.94	**Cys**	UGC*	45	1.15	2633	0.79
Leu	UUA*	304	3.64	18877	2.53	**Ser**	UCA*	182	2.90	11554	2.26	**Ter**	UAA	50	2.94	1460	2.54	**Ter**	UGA	0	0.00	134	0.23
Leu	UUG	66	0.79	5553	0.74	**Ser**	UCG	4	0.06	2102	0.41	**Ter**	UAG	1	0.06	130	0.23	**Trp**	UGG	44	1.00	3186	1.00
Leu	CUU	58	0.69	12587	1.69	**Pro**	CCU*	143	1.89	5691	1.37	**His**	CAU	23	0.34	3105	0.84	**Arg**	CGU	1	0.01	799	0.32
Leu	CUC	53	0.63	3383	0.45	**Pro**	CCC	4	0.05	1465	0.35	**His**	CAC*	114	1.66	4255	1.16	**Arg**	CGC	0	0.00	251	0.10
Leu	CUA	13	0.16	2426	0.33	**Pro**	CCA	150	1.98	7671	1.84	**Gln**	CAA*	133	1.42	4631	1.01	**Arg**	CGA	5	0.06	1338	0.53
Leu	CUG	7	0.08	1939	0.26	**Pro**	CCG	6	0.08	1840	0.44	**Gln**	CAG	54	0.58	4546	0.99	**Arg**	CGG	1	0.01	641	0.25
Ile	AUU	225	1.40	23812	1.55	**Thr**	ACU	87	1.00	8311	1.36	**Asn**	AAU	51	0.36	17956	1.30	**Ser**	AGU	54	0.86	5047	0.99
Ile	AUC*	195	1.21	8136	0.53	**Thr**	ACC	63	0.72	3585	0.59	**Asn**	AAC*	229	1.64	9706	0.70	**Ser**	AGC*	67	1.07	3383	0.66
Ile	AUA	62	0.39	14168	0.92	**Thr**	ACA*	189	2.17	9532	1.57	**Lys**	AAA*	803	1.90	38675	1.78	**Arg**	AGA*	440	5.45	8771	3.47
Met	AUG	207	1.00	12597	1.00	**Thr**	ACG	9	0.10	2934	0.48	**Lys**	AAG	43	0.10	4709	0.22	**Arg**	AGG	37	0.46	3376	1.33
Val	GUU*	361	2.33	17585	2.05	**Ala**	GCU*	293	1.82	6944	0.98	**Asp**	GAU	164	1.15	19381	1.41	**Gly**	GGU*	227	1.50	8138	0.98
Val	GUC	25	0.16	2199	0.26	**Ala**	GCC	12	0.07	1714	0.24	**Asp**	GAC*	122	0.85	8095	0.59	**Gly**	GGC	76	0.50	4103	0.50
Val	GUA	208	1.34	12102	1.41	**Ala**	GCA	322	2.00	17607	2.47	**Glu**	GAA	514	1.87	36441	1.83	**Gly**	GGA	285	1.89	17312	2.09
Val	GUG	26	0.17	2439	0.28	**Ala**	GCG	16	0.10	2221	0.31	**Glu**	GAG	36	0.13	3438	0.17	**Gly**	GGG	16	0.11	3571	0.43

Table 3-1 Codon usage in *M. maripaludis*

Summed codon usage and RSCU values for 51 highly expressed (High) genes and for all of the 1716 genes genome-wide (All). Asterisks mark codons occurring at significantly higher frequencies among highly expressed genes.

3.3 Results

3.3.1 Variation in codon usage among genes in *M. maripaludis*

Codon usage among the genes of *M. maripaludis* is biased, with an average GC3s of 0.23 among genes and A and T ending codons used at the highest frequencies (Table 3-1). There is variation in patterns of codon usage among the 1716 genes, with values of GC3s varying from 0.05 to 0.39 and N_c values ranging from 23.3 - 61. To explore this variation, the plot of N_c against GC3s (WRIGHT 1990) was examined (Figure 3-1). Whilst the range of GC3s values is narrow, with the standard deviation of GC3s among genes (0.04) only a little greater than that expected from a random distribution (0.03), values of N_c vary considerably among genes (Figure 3-1). Ribosomal protein genes have values near the lower end of the N_c distribution, indicating more biased codon usage.

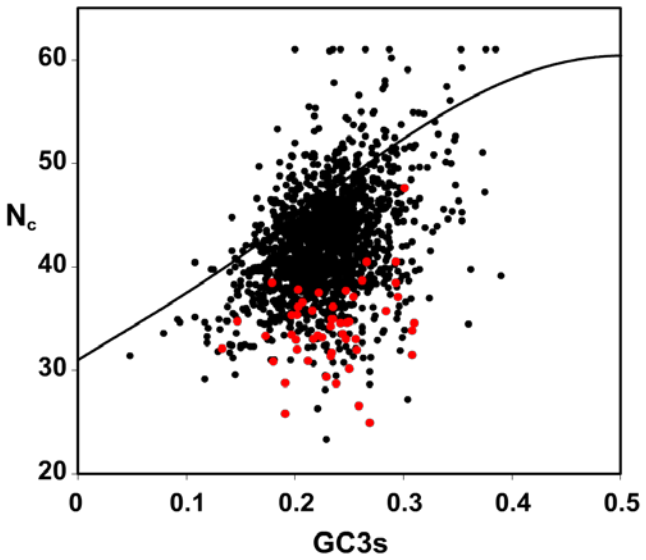


Figure 3-1 The effective number of codons (N_c) in *M. maripaludis*

Plotted against the G+C content at synonymously variable third codon positions (GC3s) (WRIGHT 1990). Red points indicate the 51 highly expressed genes. The line indicates the expected N_c with random codon usage.

Within-group correspondence analysis (WCA) revealed a major trend explaining 15.9% of variation among genes, which was found to be correlated with expression level ($r = 0.62$, $p < 0.001$). Both highly expressed genes and genes for which expression data were available fall towards the right most extreme of this axis (Figure 3-2, A&B). Other genes involved in nitrogen fixation and methane metabolism also occur at the same extreme of the primary axis, indicating similar patterns of codon bias. At the opposite end of the distribution were genes annotated as hypothetical proteins and those for which expression data were unavailable (Figure 3-2 C). The second and third WCA axes explained a small fraction of the variation in codon bias (4.5% and 4% respectively). The position of genes on the second axis did not appear to correlate with any simple index of codon usage whereas the position of genes on the third axis correlated with GC3s ($r = 0.61$, $p < 0.001$).

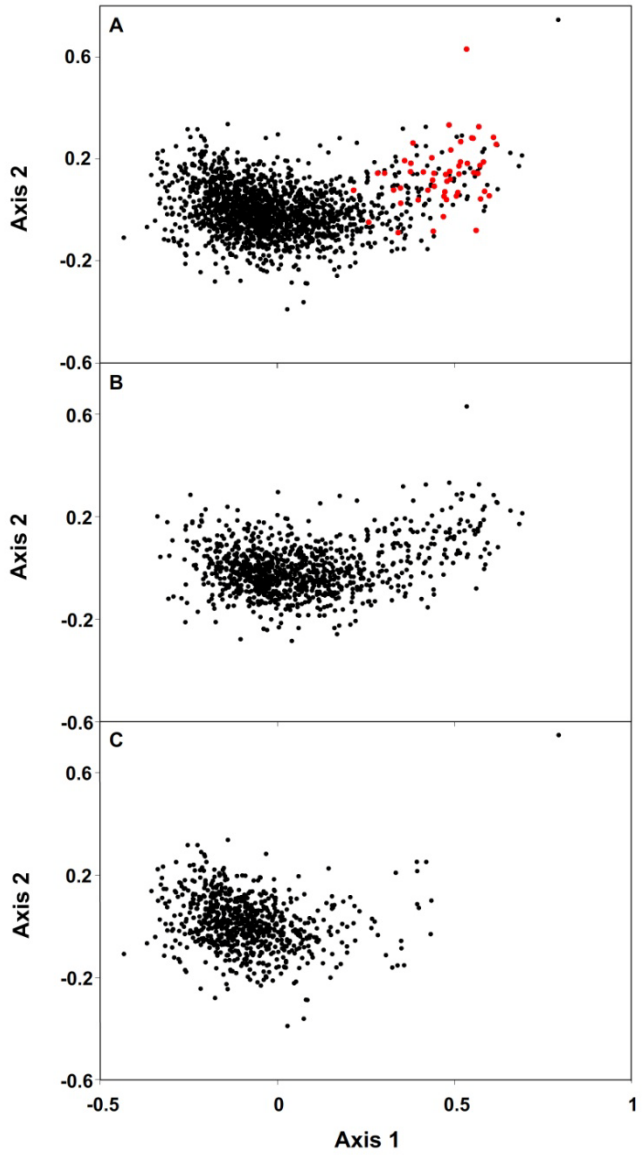


Figure 3-2 The primary and secondary axes from within block correspondence analysis in *M. maripaludis*

(A) Black points indicate all 1716 genes. Red points indicate 51 highly expressed genes. (B) Indicates the 967 genes for which expression data were available. (C) Indicates the 749 genes for which expression data were unavailable.

Eighteen codons were significantly overrepresented among highly expressed genes compared with genome wide (Table 1) and were identified as putatively optimal. These included four codons that appear to be universally optimal (Phe UUC, Tyr UAC, Ile AUC

and Asn AAC) as well as codons perfectly complementary to cognate tRNA anticodons among two-fold degenerate groups. Based upon these 18 putatively optimal codons, the frequency of optimal codons (F_{op}) was computed for each of the 1716 genes. As expected if the primary axis from correspondence analysis reflects selected patterns of codon bias, there is a strong correlation of F_{op} and axis 1 ($r = 0.89$, $p < 0.001$) and with experimental measures of protein abundance ($r = 0.59$, $p < 0.001$; Figure 3-3). Similarly the codon adaptation index (CAI) is correlated with protein abundance ($r = 0.59$, $p < 0.001$; Figure 3-4).

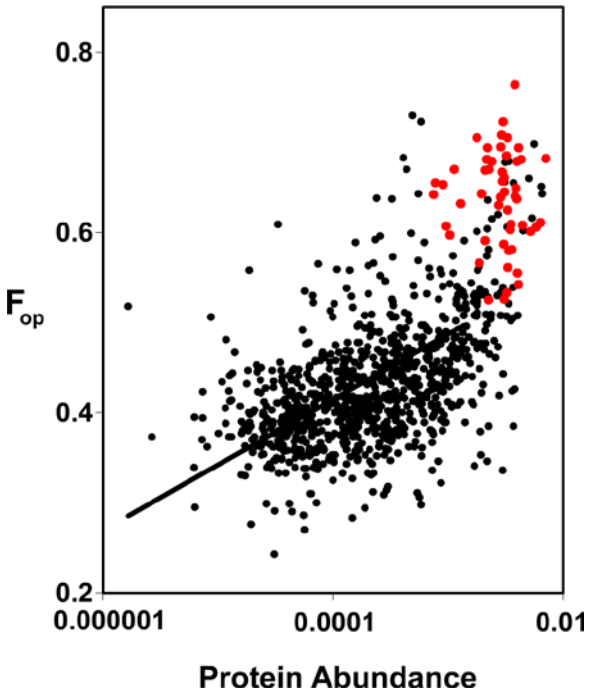


Figure 3-3 Correlation of the frequency of optimal codons (F_{op}) with protein abundance in *M. maripaludis*

Protein abundance data from Xia *et al*. (2006). Red indicates 51 highly expressed genes.

Note that protein abundance does not necessarily correspond to the most relevant component of gene expression level to select codon usage bias, i.e. the fraction of all translation time which is invested in any particular protein, because rates of protein degradation and per nucleotide rates of translation elongation may vary among genes. Similarly, if mRNA abudance were used, this would not necessarily reflect the fraction of translation time invested in each protein either, since rates of translation initiation and elongation may also vary among genes. Heterogeneity in mRNA stability among genes might also impact upon estimates of gene expression level based upon mRNA level, where mRNA levels do not reflect those at equilibrium. Finally, it is gene expression levels during periods of exponential growth, which are expected to be most relevant to most relevant to selected codon usage bias, since it is during this growth phase where most selection upon codon bias is expected to occur.

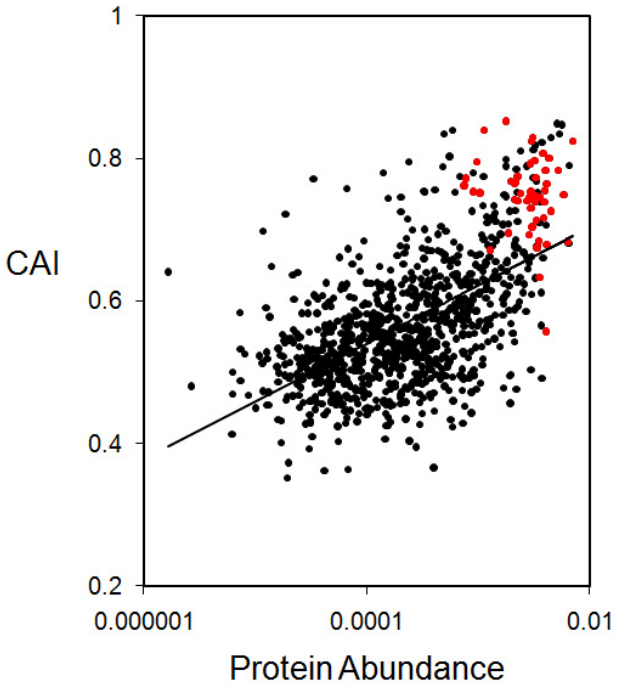


Figure 3-4 Correlation of the codon adaptation index (CAI) with protein abundance in *M. maripaludis*

Protein abundance data from Xia *et al*. (2006). Red indicates 51 highly expressed genes

3.3.2 The strength of selected codon usage bias in *M. maripaludis*

The strength of selected codon usage bias was estimated for the highly expressed genes of *M. maripaludis* and was found to be large in magnitude (S = 1.63 +0.29/-0.32), and similar to the value obtained for the archetypal example of selected codon usage bias, *Escherichia coli* (S = 1.49 +0.31/-0.29). Overall values of S are only based upon codon frequencies of two and three fold degenerate amino acids and so only inform of selection at these sites. To examine patterns of selected codon bias in more detail, the strength of selection was estimated for individual amino acids (Table 3-2), allowing for comparisons to be made among degeneracy families. Comparing the values for S among species within degeneracy families: values of S for *M. maripaludis* and *E. coli* were not significantly different at two-fold degenerate sites (p = 0.42) but values for *M. maripaludis* were significantly lower than for *E. coli* at four-fold degenerate sites ($p < 0.01$). Values of S were significantly lower at four-fold degenerate sites than at two-fold degenerate sites in *M. maripaludis* ($p = 0.03$), but not for *E. coli* ($p = 0.36$). Thus, selection appears to be weaker among four-fold degenerate sites of *M. maripaludis*.

Two-fold degenerate AAs					Four-fold degenerate AAs				
	M. maripaludis		*E. coli*		*M. maripaludis*			*E. coli*	
AA	codon	*S*	codon	*S*	AA	codon	*S*	codon	*S*
Phe	UUC	1.79	UUC	1.79	Pro	CCU	0.55	CCG	0.79
Tyr	UAC	2.27	UAC	1.39	Thr	ACA	0.61	ACU	1.52
His	CAC	1.28	CAC	1.16	Val	GUU	0.28	GUU	1.14
Asn	AAC	2.12	AAC	1.74	Ala	GCU	0.95	GCU	1.49
Asp	GAC	0.58	GAC	1.17	Gly	GGU	0.61	GGU	1.21

Table 3-2 The strength of selected codon usage bias (*S*) for two and four-fold degenerate amino acid (AA) groups in *M. maripaludis*

Only AAs for which optimal codons could be identified and tRNAs could be compared are included.

3.4 Discussion

It is clear through various lines of enquiry that natural selection is influential in shaping patterns of codon usage in *M. maripaludis*. Distinctive patterns of codon usage are observed among highly expressed genes (Table 3-1), the primary target of natural selection. These patterns are associated with a major trend describing variation in codon usage among genes; highly expressed genes fall to one extreme of the distribution (Figure 3-2) and the trend is correlated with expression level ($r = 0.62$, $p < 0.001$). Contrasting the codon usage of highly expressed genes with that used genome-wide allowed for the identification of putative optimal codons for all amino acid groups (besides Glu). Such a large number of optimal codons are associated with strong selected codon bias in Bacteria (HENRY 2007). Concordance among tRNA anticodons and optimal codons provides further indication that natural selection is present; the only available tRNA anticodons for two-fold degenerate amino acids are perfectly complementary to the identified optimal codons. Furthermore, the observed correlation of the frequency of optimal codons with expression level is only expected if natural selection shapes patterns of codon usage bias.

A previous report identified a correlation of protein abundance and Karlin's predicted highly expressed indicator and yet failed to find a correlation of protein abundance and CAI (XIA *et al.* 2006). The Karlin index (KARLIN *et al.* 2005) is not expected to be a good indicator of expression level since it (i) assumes *a priori* that selection operates upon codon usage and (ii) does not lead to the most biased genes obtaining the highest values; see (HENRY and SHARP 2007). However in this case it has performed reasonably well, probably due to the high intensity of selection in *M. maripaludis* meaning that the average codon usage among highly expressed genes is highly biased and thus a reasonable predictor of selected bias. It is not clear how CAI values were estimated in the Xia *et al*. report (2006) since they do not specify the genes used to obtain codon fitness values. Here the CAI is computed using the method of Sharp and Li (1987), with fitness values computed from codon frequencies across the 51 highly expressed genes. Contrary to the Xia *et al*. (2006) report, we observe a strong correlation of CAI and protein abundance ($r = 0.59$, $p < 0.001$; Figure 3-4). The explanatory power of this correlation is somewhat surprising as the codon adaptation index is not expected to perform well in an A+T rich genome with such strong patterns of mutational bias (GROCOCK and SHARP 2002), but here performs as well as the simpler measure of the frequency of optimal codons.

These data indicate that codon usage bias may be used as a rough approximation of expression level in *M. maripaludis*, with F_{op} and CAI both explaining 35% of variation in expression level among genes. Interestingly, genes for which no expression data were available exhibit lower F_{op} values (median = 0.39) than others (median 0.44), falling to the left of the distribution on axis 1 (Figure 3-2), as expected if these genes are expressed at low levels, atypical growth conditions or do not encode their predicted protein. There are limitations however, in using codon usage bias to approximate gene expression level. Codon usage bias is expected to best approximate the component of expression level which correponds to the fraction of all translation time that is apportioned to each protein during exponential growth (ANDERSSON and KURLAND 1990). Yet this fraction does not necessarily correspond to protein abundance level, given heterogeneity in rates of protein degradation (HOCHSTRASSER 1995; PINE 1965) and elongation rates. Second, selection is expected to be most intense upon those genes which are highly expressed during log phase, and so best predicts log phase expression level. It is therefore expected that indices of codon bias predict expression level less well with data collected under alternative growth conditions, such as the the Xia *et al*. data used in this study. Consistent with this view, the highly expressed genes selected for this study fall above the regression line in Figure 3-2, as expected if they are upregulated during exponential growth. Other reasons why selected codon usage bias does not exactly correspond to gene expression level include: sampling error and heterogeneity among genes in site specific codon preferences, such as those influencing the start of gene sequences.

A variety of well established trends describe variation in the strength of selected codon usage bias among Bacteria. Selection is strongest for species with the fastest growth rates (ROCHA 2004; SHARP *et al.* 2010). If patterns of selected codon usage bias among Archaea and Bacteria follow the same trends as a recent study indicated (VIEIRA-SILVA and ROCHA 2010), then with a growth rate which is more than six times slower than that of *E. coli*, selection in *M. maripaludis* is expected to be weak. Yet the strength of selected codon usage bias in *M. maripaludis* was found to be strong (S = 1.63 +0.29/-0.32) and not significantly different to that for *E. coli* (S = 1.49 +0.31/-0.29). It therefore remains unclear why selection is so high for *M. maripaludis,* and the possibility remains it reflects a genuine differences between Bacteria and Archaea.

The pattern of selected codon usage bias in *M. maripaludis* differs from that in *E. coli.* In *M. maripaludis* selection is weaker among four-fold degenerate sites than among two-fold degenerate sites. Whilst the reason for this difference is not entirely clear, it might relate to the distinctive tRNA requirements of each degeneracy family. A single tRNA anticodon translates the codons of two-fold degenerate amino acids and natural selection favours the best recognised codon. Each of the five four-fold degenerate amino acids are translated by a minimum of two different tRNA anticodons, and their relative abundances, which may be indicated by gene copy number (KANAYA *et al.* 1999) are relevant since natural selection typically favours the codon best recognised by the most abundant tRNA anticodon (IKEMURA 1981). The total numbers of tRNA genes in *M. maripaludis* (37) are much lower than in *E .coli* (86). Consequently, each of the four-fold degenerate amino acids in the archaeon is decoded by two different tRNA species, where each is present as single gene copy and thus expected to be expressed at equal molar ratio. The lack of major tRNA species for these amino acids might reduce the selective benefit of any single codon, since it is unlikely that any particular codon is best recognised by both different anticodons. In *E. coli*, tRNA duplication has allowed for biased patterns of tRNA usage among four-fold degenerate sites, which might better drive codon selection. This speculative interpretation remains to be tested and does not address the larger question of why tRNA duplication has not been favoured in *M. maripaludis*, and whether these patterns of selection reflect a general difference between Archaea and Bacteria.

4. TRENDS IN CODON USAGE BIAS WITHIN ARCHAEAL GENOMES

4.1 Introduction

Biased patterns of codon usage can vary among genes within species, and in Bacteria such variation has been associated with (i) heterogeneity in G+C content at synonymously variable positions and attributed to the acquisition of foreign genes, (ii) variation in the excess of G over C and/or T over A at synonymous sites, attributed to strand specific mutational biases, and (iii) variation in the usage of a subset of codons overrepresented among highly expressed genes and attributed to natural selection. Multivariate analyses have been commonly used to aid the identification of major trends in codon bias within genomes (LOBRY and CHESSEL 2003; MEDIGUE *et al.* 1991), and their application to 214 species of Bacteria has revealed widespread variation in the combination and relative importance of these trends among species (SUZUKI *et al.* 2008). In the previous Chapter, multivariate analyses were used to investigate trends in codon usage bias within a single archaeal species, *M. maripaludis*. Here, I extend these analyses to 67 species of Archaea, to determine major sources of variation in codon usage bias within archaeal genomes. By comparison with Bacteria, several differences in trends are revealed.

4.2 Methods

4.2.1 Data sources

The coding sequences for 67 species of Archaea for which complete genomes (Appendix A) were available were downloaded from GenBank using the ACNUC interface (GOUY *et al.* 1985). To reduce any sampling biases, only one representative of multiple strains of the same species was included, where arbitrarily the first genome to be sequenced was selected for analysis. Similarly, only one representative was included from closely related species sharing >95% sequence identity across 20 highly expressed genes (listed in Chapter 5), leaving 67 archaeal species in the final dataset. Where multiple chromosomes were available, only the largest chromosome was analysed, as even large second chromosomes can have atypical patterns of codon usage (HARRISON *et al.* 2010). It was not possible to use the same orthologous highly expressed genes as for Bacteria (SUZUKI *et al.* 2008) since the divergence of ribosomal protein genes means that Archaea contain alternative analogous genes to

perform the same functions (LECOMPTE *et al.* 2002). Instead a set of highly expressed genes was identified that was expected to exhibit the same average intensity of selected codon usage bias as the bacterial dataset (identified in Chapter 5).

4.2.2 Multivariate analyses

Multivariate analyses summarise high dimensional data (e.g. of 59 synonymous codons) into fewer major trends whilst retaining much of the original information (GREENACRE 2007). Of various multivariate analyses, within-block correspondence analyses (WCA) was found to be most effective at discerning trends in codon usage bias among bacterial genomes (SUZUKI *et al.* 2008), and so is employed here as the method of choice. WCA takes the codon usage summed for each of the genes within a genome and successively identifies axes explaining the most variation, whilst accounting for variation in amino acid usage. Here, WCA was implemented in the ade4 package (CHARIF *et al.* 2005) of the R suite. The resulting axes may reflect biological phenomena or random noise (SUZUKI *et al.* 2008). The source of variation that was attributed to each axis was identified using the same somewhat stringent criteria previously applied to bacterial genomes (SUZUKI *et al.* 2008), nevertheless these criteria allow direct comparisons among Archaea and Bacteria to be made. Briefly, coordinates of genes upon the first three resulting axes were correlated with (i) synonymously variable G+C content across third codon positions (GC3s) and (ii) the excess of G over C at synonymously variable third codon positions (G-C skew). Variation upon an axis was attributed to either of these factors in the instances where correlation coefficients of $r > 0.7$ were obtained. Axes attributed to variation in selected codon usage bias were identified as the instances where the mean standard score across 20 highly expressed genes was > 1.64 (SUZUKI *et al.* 2008). In further analyses to identify additional asymmetric mutational patterns, the coordinates of genes upon the three resulting axes were also correlated with the excess of T over A (T-A skew), the combined excesses of G+T over C+A (K-M skew) and of C+T over G+A (Y-R skew) at synonymously variable third codon positions. Strand skews were identified as the source of variation for axes where correlation coefficients were >0.7 and the 'major skew factor' was identified as the skew index producing the highest correlation coefficient.

4.2.3 Exploring variation among genes

Cumulative skew plots were examined to explore variation in patterns of skew with respect to the location of *cdc6* and *cdv* genes commonly associated with the location of replication origins (LINDÅS *et al.* 2008; LUNDGREN *et al.* 2004). To explore heterogeneity in G+C content at synonymously variable third codon positions (GC3s), the ratio of observed over expected standard deviation of GC3s was computed, where the expected value was estimated using binomial theory. To explore patterns of GC3s variation within genomes, the G+C content at synonymously variable third codon positions was plotted against the gene position across the chromosome.

4.3 Results

4.3.1 Trends in codon usage within archaeal genomes

Within group correspondence analysis (WCA) was implemented to detect major trends in codon usage associated with (i) G+C content at synonymously variable third codon positions (GC3s), (ii) asymmetric mutational patterns and (iii) expression level within 67 archaeal genomes. Trends were found to be associated with variation in GC3s for most (58) species, with variation in strand skew for more than half (36) of species, and in expression level for a small minority (4) of species (Table 4-1). A single factor was identified to explain heterogeneity in codon usage among genes for approximately half (33) of species, and two major factors described variation in codon usage for most of the remaining others (31). There were two species for which no factors were detected, and none for which all factors were detected although any factor may be present or absent in each species.

Each WCA axis successively explains variation among genes, with the first axis explaining the most variation. Among 67 species of Archaea, variation explained by the first axis varies from 7.5% in *Nanoarchaeum equitans* to 34.0% in *Methanocella paludicola,* with a median of 17.1%. Factors were identified upon the first major axis in the majority (63) of species, and most of these trends (54) were associated with GC3s (Table 4-1). In four species no factors were detected upon the first major axes, and for these species first axes explained a relatively small proportion (7.5% - 10%) of variation among genes. Variation explained by the second axes ranged from 3.7% in *Methanosaeta thermophila* to 15.0% in *Pyrobaculum calidfontis* with a median of 5.6%. Factors were identified on second axes for many of the species (28) and

most (25) were associated with strand skews. Third axes generally explained little variation among genes (median 4.1%) and factors for this axis were only detected in nine species.

Species	Within-block correspondence analysis[a]			GC3s[b]	SD[c]	obs/exp[d]
	Axis 1	Axis 2	Axis 3			
Euryarchaeaota						
Archaeoglobales						
Archaeoglobus fulgidus	GC3s	K-M skew	- -	0.56	0.074	1.99
Archaeoglobus profundus	GC3s	- -	K-M skew	0.43	0.064	1.63
Ferroglobus placidus	GC3s	- -	K-M skew	0.48	0.070	1.79
Halobacteriales						
Haloarcula marismortui	GC3s	- -	- -	0.75	0.094	2.84
Halobacterium salinarum	GC3s	K-M skew*	- -	0.87	0.089	3.49
Halorubrum lacusprofundi	GC3s	- -	- -	0.86	0.071	2.70
Halomicrobium mukohataei	GC3s	- -	- -	0.85	0.087	3.11
Halorhabdus utahensis	GC3s	- -	- -	0.79	0.089	2.85
Haloterrigena turkmenica	GC3s	- -	- -	0.87	0.067	2.49
Haloquadratum walsbyi	GC3s	- -	- -	0.42	0.108	2.66
Natrialba magadii	GC3s	G-C skew*	- -	0.76	0.028	1.42
Natronomonas pharaonis	GC3s	G-C skew*	T-A skew	0.77	0.072	2.26
Methanobacteriales						
Methanobacterium thermoautotrophicus	GC3s	K-M skew	- -	0.54	0.080	2.19
Methanobrevibacter ruminantium	GC3s	Expression	- -	0.30	0.076	2.28
Methanobrevibacter smithii	K-M skew*	- -	- -	0.18	0.045	1.59
Methanosphaera stadtmanae	K-M skew*	- -	- -	0.11	0.035	1.58
Methanopyrales						
Methanopyrus kandleri	GC3s	G-C skew*	- -	0.74	0.089	2.94
Methanococcales						
Methanocaldococcus jannaschii	- -	K-M skew	- -	0.22	0.042	1.40
Methanocaldococcus fervens	Expression	- -	GC3s	0.23	0.042	1.35
Methanocaldococcus vulcanius	K-M skew*	- -	GC3s	0.24	0.041	1.28
Methanococcus aeolicus	Expression	- -	- -	0.20	0.035	1.25
Methanococcus maripaludis strain S2	Expression	- -	- -	0.23	0.040	1.33
Methanococcus vannielii	Expression	- -	- -	0.20	0.041	1.45
Methanosarcinales						
Methanococcoides burtonii	GC3s	K-M skew	- -	0.37	0.074	2.16
Methanosaeta thermophila PT	GC3s	- -	- -	0.64	0.065	1.92
Methanosarcina acetivorans	GC3s	- -	- -	0.44	0.092	2.58
Methanosarcina barkeri fusaro	GC3s	- -	- -	0.37	0.075	2.16
Methanosarcina mazei	GC3s	- -	- -	0.41	0.080	2.24
Rice Cluster I MRE50	GC3s	- -	- -	0.70	0.115	3.57
Methanomicrobiales						
Methanocorpusculum labreanum	GC3s	G-C skew*	- -	0.56	0.093	2.72
Methanoculleus marisnigri JR1	GC3s	- -	K-M skew	0.80	0.098	3.34
Methanoregula boonei 6A8	GC3s	- -	- -	0.64	0.089	2.61
Methanospirillum hungatei	GC3s	- -	K-M skew	0.43	0.083	2.42
Methanocella paludicola	GC3s	K-M skew	- -	0.74	0.125	3.84
Methanosphaerula palustris	GC3s	- -	- -	0.67	0.113	3.13

Species	Within-block correspondence analysis[a]			GC3s[b]	SD[c]	obs/exp[d]
	Axis 1	Axis 2	Axis 3			
Euryarchaeaota						
Thermococcales						
Pyrococcus abyssi	GC3s	- -	- -	0.49	0.073	2.13
Pyrococcus furiosus	GC3s	- -	K-M skew	0.37	0.077	2.16
Pyrococcus horikoshii	- -	GC3s	- -	0.41	0.070	2.00
Thermococcus gammatolerans	GC3s	- -	- -	0.68	0.086	2.51
Thermococcus kodakarensis	GC3s	- -	K-M skew	0.64	0.093	2.61
Thermococcus onnurineus	GC3s	- -	- -	0.63	0.085	2.35
Thermococcus sibiricus	GC3s	K-M skew*	- -	0.36	0.079	2.07
Thermoplasmatales						
Picrophilus torridus	GC3s	- -	- -	0.34	0.062	1.88
Thermoplasma acidophilum	GC3s	- -	- -	0.52	0.079	2.31
Thermoplasma volcanium	GC3s	- -	- -	0.39	0.062	1.78
Aciduliprofundum boonei	GC3s	K-M skew	- -	0.35	0.062	1.41
Crenarchaeota						
Desulfurococcales						
Aeropyrum pernix	GC3s	K-M skew	- -	0.66	0.106	3.10
Hyperthermus butylicus	GC3s	K-M skew	- -	0.57	0.086	2.43
Ignicoccus hospitalis	GC3s	- -	- -	0.75	0.071	2.26
Staphylothermus marinus	K-M skew*	GC3s	- -	0.26	0.047	1.48
Desulfurococcus kamchatkensis	GC3s	T-A skew	- -	0.47	0.079	1.96
Thermoproteales						
Caldivirga maquilingensis	K-M skew*	- -	- -	0.43	0.057	1.67
Pyrobaculum aerophilum	GC3s	K-M skew	- -	0.58	0.088	2.17
Pyrobaculum arsenaticum	GC3s	K-M skew*	- -	0.66	0.102	2.88
Pyrobaculum calidifontis	GC3s	K-M skew*	- -	0.70	0.094	2.82
Pyrobaculum islandicum	GC3s	K-M skew*	- -	0.51	0.146	3.93
Thermoproteus neutrophilus	GC3s	K-M skew*	- -	0.76	0.085	2.66
Thermofilum pendens	GC3s	K-M skew	- -	0.72	0.101	3.17
Sulfolobales						
Metallosphaera sedula	GC3s	- -	- -	0.51	0.083	2.31
Sulfolobus acidocaldarius	GC3s	K-M skew	- -	0.31	0.073	2.17
Sulfolobus islandicus	GC3s	K-M skew	- -	0.28	0.060	1.49
Sulfolobus solfataricus	GC3s	K-M skew	- -	0.31	0.103	2.58
Sulfolobus tokodaii	GC3s	K-M skew	- -	0.23	0.083	2.62
Thaumarchaeota						
Nitrosopumilus maritimus	- -	- -	- -	0.20	0.052	1.73
Cenarchaeum symbiosum A	GC3s	- -	- -	0.68	0.109	2.92
Korarchaeota						
Korarchaeum cryptofilum	GC3s	- -	- -	0.57	0.060	1.73
Nanoarchaeota						
Nanoarchaeum equitans	- -	- -	- -	0.23	0.030	1.00

Table 4-1 Major trends in codon usage within 67 genomes of Archaea

[a]Factors associated with WCA axes
[b]Average G+C content across synonymously variable third codon positions (GC3s)
[c]The standard deviation of GC3s among genes
[d]The ratio of observed to expected standard deviation of GC3s
* Indicates where skew detected under G-C skew Suzuki criterion

Trends were detected using criteria previously applied to bacterial genomes (SUZUKI *et al.* 2008) to allow for direct comparisons to be made among Archaea and Bacteria. However, these criteria might lead to detection issues. First, using a correlation coefficient threshold of $r > 0.7$ might systematically bias factors detected by elevating the relative importance of secondary and tertiary trends relative to primary trends. For instance a factor correlated with a primary axis explaining 17% of variation among genes with a coefficient as low as $r = 0.5$ would explain more variation in codon usage than a factor correlated with a secondary axis explaining 7% of variation with a coefficient of $r = 0.7$. Among 67 archaeal genomes, this did not appear to be a major problem. There were only two species, *Methanocaldococcus jannaschii* and *Pyrococcus horikoshii*, for which factors were detected upon second but not first axes, and in both cases second axes (Metjan 7%; Pyrhor 9%) explained almost as much variation as first axes (Metjan 9%; Pyrhor 10%) meaning that the factors strongly associated with second axes were indeed more important than factors weakly associated with first axes. Second, using threshold criteria to detect factors means that factors occuring just below the threshold can be overlooked. There are at least two instances where this has occurred among Archaea. For *Pyrococcus horikoshii*, the primary axis is correlated with an index of strand skew (CT - GA) with a coefficient of $r = 0.62$ and a cumulative skew plot reveals a typical replication associated pattern consistent with skew as a major factor contributing to heterogeneity in codon usage. The same is the case for the second axis of *Haloquadratum walsbyi*, which is correlated with G-C skew ($r = 0.63$, $p < 0.01$). These instances where WCA has failed to detect trends are generally of smaller effect and explain less variation in codon usage than trends which were successfully detected, therefore overlooking trends may not be a large problem when the objective is to detect major trends of large effect.

4.3.2 Comparing trends among Archaea and Bacteria

Trends describing patterns of codon usage among genes were detected for 67 species of Archaea, and here are compared with trends previously detected for 214 species of Bacteria (SUZUKI *et al.* 2008). Among both domains, the factor most commonly identified as a source of variation among genes was GC3s, which produced major trends in codon usage within a similar proportion of archaeal (87%) and bacterial (89%) genomes. Trends in GC3s are typically large in magnitude, and dominate the first WCA axes of both domains (Table 4-2). Trends associated with G-C skew were underrepresented among Archaea, occurring in 22% of species compared with half of bacterial species ($\chi^2 = 15.8$; $df = 1$; $p < 0.01$), however a much

larger proportion of Archaea (55%) exhibited major trends associated with skew of some form. Trends in G-C skew are commonly of secondary importance, with most occurring upon the secondary WCA axis in species of both domains. Trends associated with gene expression level were highly underrepresented among Archaea, occurring in only 7% of species compared with two thirds of bacterial species ($\chi^2 = 33.8$; $p < 0.001$).

Factor	Axis 1				Axis 2				Axis 3			
	Archaea		Bacteria		Archaea		Bacteria		Archaea		Bacteria	
	N	%	N	%	N	%	N	%	N	%	N	%
GC3s	54	80.6	150	70.1	2	3.0	34	15.9	2	3.0	7	3.3
G-C skew	5	7.5	38	17.8	10	14.9	57	26.6	0	0.0	13	6.1
All skew	5	7.5	--	--	25	37.3	--	--	7	10.4	--	--
Expression	4	6.0	37	17.3	1	1.5	52	24.3	0	0.0	53	24.8

Table 4-2 Summary of factors associated with variation in codon usage within archaeal and bacterial genomes

The number (N) and percentage (%) of species among Archaea and Bacteria for which factors (shown) were associated with within-group correspondence analysis axes. G-C skew refers only to skews detected using the method of Suzuki *et al*. (2008).

4.3.3 Trends in strand skew within the genomes of Archaea

Trends in strand skew were identified in 36 species of Archaea. The method previously applied to bacterial genomes (SUZUKI *et al.* 2008) only identified skew trends based upon the cumulative excess of G over C, and under these criteria, only 15 of the 37 trends among Archaea are detected (Table 4-2). Trends in G-C skew were only the most important skew factor (i.e. detected rather than K-M) in four species. Similarly, the cumulative excess of T over A was only the most important skew factor in a single species, *Desulfurococcus kamchatkensis,* although clearly both G-C and T-A skews contribute to the numerous K-M skews observed among Archaea. Whilst most skew patterns reflect heterogeneity in K-M among genes, there is some variation in the direction of strand skews among Archaea. For instance, in *Natronomonas pharaonis*, skew patterns are observed for both G-C and T-A but this does not reflect heterogeneity in K-M among genes. The cumulative excess of G over C and T over A are remarkably uncorrelated ($r = 0.08$), have opposing directions (Figure 4-1) and were detected on distinct orthogonal axes. Cumulative skew plots were examined

among Archaea with respect to the location of replication-origin-associated *cdc6* and *cdv* genes (LINDÅS *et al.* 2008; LUNDGREN *et al.* 2004) and in many instances abrupt sign changes coincided with the locations of these genes (e.g. Figure 4-1).

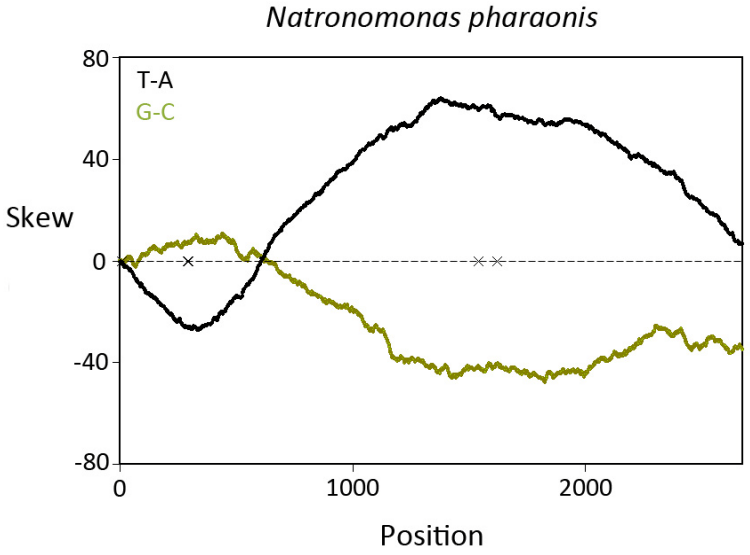


Figure 4-1 Cumulative skew plot for *Natronomonas pharaonis*

Cumulative sum of G-C and T-A skew values for genes across the chromosome of *Natronomonas pharaonis*. The arbitrary start position is that of the first annotated gene. The location of replication-origin-associated *cdc6* genes are indicated with crosses.

4.3.4 Trends in GC3s within the genomes of Archaea

Trends in GC3s were identified for 58 species of Archaea. Heterogeneity in GC3s is expected to follow a binomial distribution with the greatest variation expected at intermediate base composition. To explore variation in GC3s independently of expected binomial variation in Archaea, the ratio of observed to expected standard deviation of GC3s (SD GC3s) was estimated (Table 4-1), and is expected to be equal to 1 in the absence of region specific mutational biases, natural selection or lateral transfer events. In *N. equitans*, where WCA identified no trends in codon usage, the observed to expected ratio of SD GC3s was equal to 1.00, indicating no sources of additional bias. Aside from the case of *N. equitans*, the observed to expected SD GC3s varies among Archaea from 1.25 in *Methanococcus aeolicus* to 3.94 in *Pyrobaculum islandicum*, with an average value of 2.28. The observed to expected SD GC3s was found to be highly correlated with genome-wide GC3s ($r = 0.68$, $p < 0.001$; Figure 4-2).

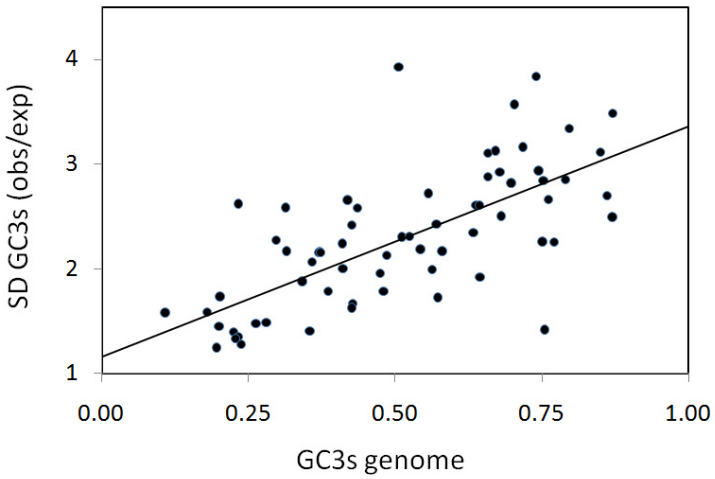


Figure 4-2 Correlation of the ratio of observed to expected standard deviation of GC3s with genome-wide GC3s

Among bacterial genomes, horizontally transferred genes giving rise to variation in GC3s often occur in clusters upon a chromosome in genomic islands. Plots of GC3s against gene position upon the chromosome were explored to give an indication of atypical codon usage commonly associated with lateral transfer events. Several instances of apparent large genomic islands (~20-60 kb regions of atypical GC3s) common among Bacteria were observed in the genomes of Archaea. One example is a region from gene position 434-498 in *Sulfolobus acidocaldarius* (Figure 4-3), which contains a mixture of hypothetical proteins, conjugative proteins, and transposases. By far the greatest source of variation in GC3s in *S. acidocaldarius* was due to regional variation in GC3s across the chromosome. In the *S. acidocaldarius* genome, GC3s was found to be correlated with cumulative K-M skew among genes ($r = 0.53$, $p < 0.001$), and sign changes were found to coincide with the location of experimentally determined replication origins, as determined experimentally by Lundgren *et al.* (2004). Further instances of regional variation in GC3s were observed across the chromosomes of other Archaea (Figure 4-3).

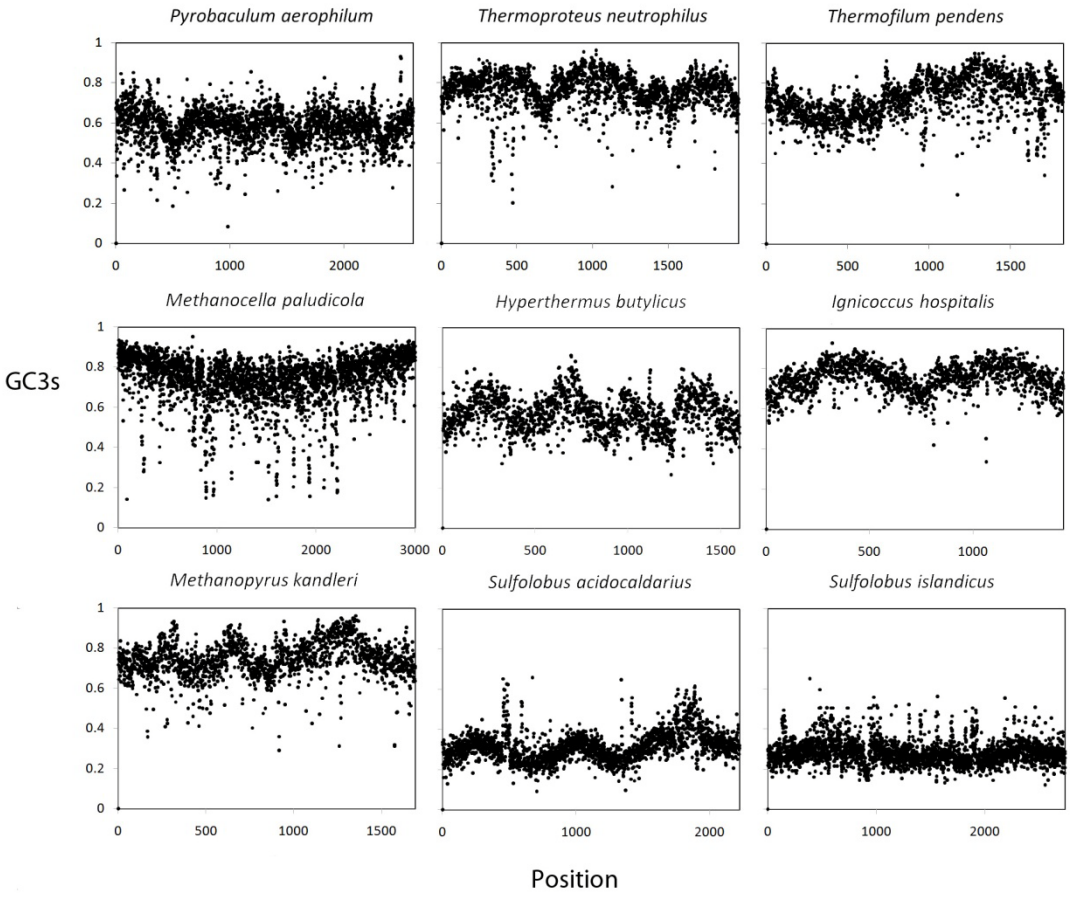


Figure 4-3 Patterns of GC3s across the chromosomes of nine species of Archaea

The G+C content at synonymously variable third codon positions (GC3s) for each protein coding gene against its relative chromosomal position. The start position is arbitrary, relative to the first gene in the annotated genome sequence. The nine species were selected on the basis that their plots exhibit the most varied and visually striking patterns. Some of these species (*P. aerophilum*, *M. paludicola* and *M. kandleri*) were selected because they have large standard deviations of GC3s values. *T. pendens* was selected because it exhibits a single phase pattern reminiscent of a bacterial cumulative skew plot. *T. neutrophilus* was selected because it belongs to the same family as *T. pendens*, and yet its pattern of GC3s differs. *I. hospitalis* was chosen because it exhibits a two-phase pattern of GC3s which is not present in other species in its family such as *H. butylicus*. Finally, *S. acidocaldaius* was chosen because it exhibits a three-phase of GC3s which is not present in its closely related sister taxon *S. islandicus.*

4.4 Discussion

Within-block correspondence analysis (WCA) has revealed major trends in codon usage within the genomes of Archaea. Trends were associated with the factors (i) GC3s (ii) strand skew and (iii) gene expression level, as previously shown to explain variation in codon

usage within bacterial genomes (SUZUKI *et al.* 2008). These three factors appear to succinctly explain major trends in variation within archaeal genomes since WCA axes for which no factor was determined explain less variation in codon usage among genes. The study of codon usage within bacterial genomes concluded that each of the three factors may be present or absent in each species, and their relative importance varies. Whilst the same appears to be true here for species of Archaea, there are also some more subtle patterns. Among both domains, trends associated with G+C content dominate the primary axis, indicating heterogeneity in G+C content is typically the most important source of variation in codon usage among genes. Similarly, among Archaea trends associated with strand skew are most commonly associated with the second WCA axis, indicating that skew is typically the second most important factor.

4.4.1 The causes of strand skew in Archaea

Among bacterial genomes, strand skews are commonly associated with DNA replication, such that cumulative skew plots abruptly change sign at the origin and terminus of replication (LOBRY 1996; MCLEAN *et al.* 1998). Strand skews among Archaea also appear to be replication-associated; a cumulative plot of K-M skew in *S. acidocaldarius* was found to change sign at the location of experimentally determined replication origins. Similarly, cumulative skew plots for other species of Archaea change sign at locations that coincide with the positions of replication-origin-associated *cdc6* and *cdv* genes (e.g. Figure 4-1). The occurrence of multiple origins of replication impacts substantially upon skew patterns among Archaea. In *S. acidocaldarius* where three replication origins have been determined, skew patterns exhibit periodicity with three maxima and minima, which are highly correlated with the GC3s plot (Figure 4-3).

In most bacterial species, the leading strand of DNA replication is enriched in G over C, and to a lesser extent T over A, meaning that it is G+T rich (FRANCINO and OCHMAN 1997; LOBRY 1996). In some bacterial species, such as *Mycoplasma genitalium,* the T-A skew can be stronger than the G-C skew (MCLEAN *et al.* 1998), and so for these species, skew patterns are unlikely to be detected on the basis of G-C skew alone. Indeed a previous study which only examined G-C skew failed to identify any skew associated trend in *M. genitalium* (SUZUKI *et al.* 2008), and so may have overlooked skew-associated trends in other Bacteria. In these analyses of

archaeal genomes, the identification of skew trends on the basis of G-C alone failed to identify 60% of all skew associated trends, indicating an important role for T-A and combined K-M skews among Archaea. A larger proportion of bacterial than archaeal genomes were found to exhibit G-C skews and it remains unclear whether this difference reflects a dominance of T-A skews, or a lower proportion of skew-associated trends in Archaea than in Bacteria. This question could be addressed by the identification of T-A and K-M skews within bacterial genomes.

Whilst the leading strand of DNA replication is typically G+T rich among bacterial genomes, there is some variation in the direction of strand skew. For instance, in the bacterium *Bacillus anthracis*, the T-A skew is inverted such that the leading strand is G+A rich (NECSULEA and LOBRY 2007). There is also variation in the direction of strand skews in Archaea. An experimental study which determined the replication origin for *Halobacterium sp*. (excluded from these analyses due to its close relatedness to *Halobacterium salinarum*), indicates that the leading strand is enriched in C over G (BERQUIST and DAS SARMA 2003; WORNING *et al.* 2006). Here, opposing directions of the cumulative G-C and T-A strand skews in *N. pharaonis* indicate variation in skew direction, and the concordance of skew patterns with replication origin-associated genes presumably indicates that they reflect DNA replication-associated mutational biases. In this case however, it is not possible to discern whether the leading strand is G+A or C+T rich since replication origin-associated genes occur at both the skew maxima and minima (Figure 4-1).

4.4.2 The causes of variation in GC3s among genes in Archaea

Among both Archaea and Bacteria the first WCA axis is most commonly explained by heterogeneity in GC3s, and there was no significant difference observed in the proportion of either domain exhibiting trends in GC3s. Within bacterial genomes, heterogeneity in GC3s is typically attributed to the presence of horizontally transferred genes (GROCOCK and SHARP 2002; MEDIGUE *et al.* 1991; OCHMAN *et al.* 2000), and here there seems to be substantial evidence for a widespread role of horizontal gene transfer among the genomes of Archaea. Horizontally transferred genes in bacterial genomes can often occur in clusters known as islands. Here within archaeal genomes, patterns of atypical GC3s often occur in clusters across the chromosome (Figure 4-3) and contain genes indicative of mobile elements,

consistent with lateral transfer. Horizontally transferred genes are typically A+T rich (DAUBIN *et al.* 2003; OCHMAN *et al.* 2000), as are the plasmids which may distribute them (VAN PASSEL *et al.* 2006), and so the presence of acquired genes is expected to contribute more substantially to heterogeneity in GC3s among G+C rich, rather than A+T rich species. Consistent with this, among Archaea, genome-wide GC3s is positively correlated with the ratio of observed to expected standard deviation of GC3s (Figure 4-2).

Not all heterogeneity in GC3s among Archaea can be attributed to variation in horizontally transferred genes. In a large number of species, systematic patterns of GC3s across the chromosome were observed (Figure 4-3). Such patterns have not been observed among bacterial genomes (DAUBIN and PERRIERE 2003) with the possible exception of *M. genitalium* (KERR *et al.* 1997), and may indicate a difference between Archaea and Bacteria. The correlation of GC3s with cumulative K-M skew among the genes of *S. acidocaldarius* and their association with the location of experimentally determined replication origins is consistent with a replication-associated mutational bias as the underlying molecular cause. Nevertheless the considerable variation in GC3s patterns within the genomes of Archaea may indicate a variety of causes, which will be interesting to investigate in the future.

4.4.3 Conclusions

Trends in codon usage within the genomes of Archaea appear to share some common and some different features with trends identified among Bacteria. Both exhibit replication-associated patterns of strand skew, and yet only those among Archaea exhibit multiple periodicity. Both exhibit heterogeneity in GC3s due to horizontally transferred genes; yet regional heterogeneity in GC3s appears to be largely restricted to the archaeal domain. The observation that trends associated with gene expression are dramatically underrepresented among Archaea might indicate that selected codon usage bias is a less common feature of archaeal genomes, and this will form the subject of the next chapter.

5. THE STRENGTH OF SELECTED CODON USAGE BIAS IN ARCHAEA

5.1 Introduction

The strength of selected codon usage bias (*S*) indicates the impact natural selection has had in shaping the patterns of synonymous codon usage of an organism. For species where selection is effective, it is the highly expressed genes that display the greatest bias towards a subset of optimal codons best recognised by the most abundant tRNA species (BENNETZEN and HALL 1982; GRANTHAM *et al.* 1981; IKEMURA 1981), for the efficiency and/or accuracy of translation. Therefore the strength of selected codon usage bias can be gauged by contrasting patterns of codon usage in highly expressed genes with other genes. This approach has been taken to estimate *S* for a wide variety of bacterial and eukaryotic species (DOS REIS and WERNISCH 2008; SHARP *et al.* 2005) but as yet, has not been applied to the Archaea.

Across bacterial genomes, the strength of selected codon usage bias varies with life history strategies (ROCHA 2004; SHARP *et al.* 2005; SHARP *et al.* 2010), with values of *S* largest among free-living fast-growing species with large numbers of rRNA operons and tRNA genes. Archaea have low numbers of rRNA operons and tRNA genes by comparison with bacterial species. Thus, if they conform to the trends observed across Bacteria, then they will show little or no evidence for selected codon usage bias. There has been some indication so far in these investigations that this is the case, since in the preceding chapter, natural selection was found to shape patterns of codon usage in proportionally fewer species of Archaea than Bacteria. To investigate variation in selected codon usage bias across Archaea, here the strength of this selection is estimated for 67 species.

Among Archaea, selection was found to have had less impact in shaping patterns of codon usage by comparison with some bacterial species. A linear model revealed that only two factors, minimal generation time and optimal growth temperature, explain significant variation in *S*. Across bacterial genomes by comparison, the numbers of rRNA operons, tRNA genes and growth rate all contribute to variation in *S* (SHARP *et al.* 2005; SHARP *et al.* 2010; VIEIRA-SILVA and ROCHA 2010). Species living at high temperatures exhibited systematically lower values of *S* given their growth rate, and evidence is presented for why this reflects a kinetic impact of temperature upon translation elongation. The implications of this for the nature of translational selection are discussed.

5.2 Methods

5.2.1 Data sources

The same dataset of coding and complete genome sequences for each of 67 species of Archaea was used as in Chapter 4 (Appendix A). Numbers of rRNA operons and protein coding genes, and chromosome sizes were obtained from genome annotation. Numbers of tRNA genes were obtained from the tRNA scan search server database (LOWE and EDDY 1997). Maximal growth rates and optimal growth temperatures were obtained from the literature (Appendix B).

5.2.2 Estimating *S*

The strength of selected codon usage bias (*S*) was estimated for each species by adapting a method previously developed to study bacterial genomes (SHARP *et al.* 2005), whereby codon usage among highly expressed genes is contrasted with genome-wide codon usage. Genome-wide codon usage was summed across all protein coding genes for the four amino acids phenylalanine (Phe), isoleucine (Ile), tyrosine (Tyr) and asparagine (Asp). Whilst the use of genome-wide codon frequencies inevitabily contains a small fraction of genes under selected codon usage bias, thus downwardly biasing estimates of *S*; it seems that their usage provides a more reliable indicator of neutral codon frequencies than those of lowly expressed genes, because these genes are often subject to more frequent lateral transfer, and exhibit atypical patterns of codon usage.

It was not possible to use the same orthologous set of highly expressed genes as were used in a study of bacterial genomes (SHARP *et al.* 2005). To obtain estimates of *S* among Archaea which could be reliably compared in magnitude with values among Bacteria, *S* values were first estimated based upon two highly expressed translation elongation factor genes conserved among Archaea and Bacteria (*tufA* and *fusA*), for each of the 67 species of Archaea analysed here, and the 80 species of Bacteria previously analysed (SHARP *et al.* 2005). Based upon patterns of codon usage bias, these elongation factor genes appear to be substantially more highly expressed than any other highly expressed genes. Across 80 bacterial species, the strength of selected codon bias across elongation factor genes (S_{EF}) was found to be around 30% higher than *S* based upon the 40 highly expressed genes previously analysed (SHARP *et al.* 2005). A highly expressed gene set for Archaea was therefore determined that

(i) contained genes exhibiting S values ~30% lower than archaeal S_{EF} values, and (ii) contained ~1000 Phe, Ile, Tyr and Asn codons in each species. The final dataset contained the 20 genes: *tufA*, *fusA*, *atpA*, *atpB*, *rpoA*, *rpoB*, *rpl1*, *rpl3*, *rpl10*, *rpl10e*, *rpl12*, *rpl18*, *rpl24e*, *rps2*, *rps3*, *rps4*, *rps5*, *rps7*, *rps8* and *rps15*. Orthologues for these genes were identified using a combination of genome annotation and BLAST searches (ALTSCHUL *et al.* 1990), and were manually verified by inspection of Clustalw (THOMPSON *et al.* 1994) alignments to ensure that no gene regions were missing. Confidence intervals were determined by permutation (see Chapter 2).

5.2.3 Phylogenetic analyses

Genome annotation was used to identify 80 genes expected to be highly conserved on the basis of their functions as ribosomal protein genes, translation elongation factors, ATPases or RNA polymerases. Of these, only 45 genes were found to be conserved among all Archaea, and thus suitable for phylogenetic analysis. The dataset was further restricted to those genes which showed no obvious evidence of horizontal gene transfer. Putative transfer events were assessed by comparing individual gene trees constructed by the neighbour joining method (SAITOU and NEI 1987) with the consensus tree. Any discordance of family-level clade topology, or instances where species were assigned to an icorrect family with more than 80% bootstrap support, were considered to be putative gene transfer events, leaving fourteen genes in the final dataset: *tufA*, *fusA*, *rpoA*, *rpoB*, *atpA*, *atpB*, *rpl1*, *rpl2*, *rpl3*, *rpl4*, *rps2*, *rps3*, *rps4*, and *rps5*. Note that the above criteria are conservative; this does not indicate that 31/45 genes have been laterally transferred. Protein sequences for these genes were aligned in Clustalw (THOMPSON *et al.* 1994) and any ambiguous regions or sites containing gaps were removed. The phylogeny of Archaea was estimated by maximum likelihood (ML) under the WAG model of protein evolution (WHELAN and GOLDMAN 2001) with gamma distributed rates among sites as implemented in PhyML (GUINDON and GASCUEL 2003).

To assess trends among Archaea whilst controlling for shared ancestry, standardised (GARLAND *et al.* 1992) phylogenetic independent contrasts (FELSENSTEIN 1985) were obtained using the ape (PARADIS *et al.* 2004) package of the R suite. Here, the only correlation coefficients reported are of independent contrasts since others may be confounded by the underlying phylogeny.

5.3 Results

5.3.1 The strength of selected codon usage bias in Archaea

The strength of selected codon usage bias (*S*), which indicates the degree to which natural selection has been influential in shaping patterns of codon usage across highly expressed genes, was estimated for each of 67 species of Archaea. Values of *S* vary among species from -0.275 in *Thermofilum pendens* to 1.805 in *Methanococcus maripaludis* (Table 5-1). For 28% of species including *T. pendens*, values of *S* were not significantly different from zero, indicating that selection is ineffective in these species. This is greater than the number of species in Chapter 4 for which WCA (within block correspondence analysis) axis were attributed to natural selection. This discrepancy is likely to reflect the relative insensitivity of WCA at detecting small effects, since many Archaea have $S < 1$, indicting weak selection. Values of *S* across the 67 species of Archaea in these analyses (mean 0.60) are somewhat lower than values among 80 bacterial species (mean 0.75) analysed in a previous study (SHARP *et al.* 2005), largely due to a deficit of archaeal species exhibiting very strong selected codon usage bias. The strongest selected codon usage bias among Archaea is observed among the family of the anaerobic inhabitants of marine marsh sediment, the Methanococcales (Table 5-1) where values of *S* range from 0.49 to 1.81. Across all Archaea only two species (3%) have values of $S > 1.5$, however these values are not as high as those observed for many Bacteria, where 13/80 (16%) of species have *S* values >1.5, with the highest value of 2.65 for *Clostridium perfringens* (SHARP *et al.* 2005).

Species	Gene Numbers[b]			GenT[c]	Temp[d]	GC content[e]		S[f]	Random[g]
	rRNA	tRNA	ORFs			(i)	(ii)		
Euryarchaeaota									
Archaeoglobales									
Archaeoglobus fulgidus	1	47	2436	4.0	83	0.49	0.56	0.382*	(0.214/-0.231)
Archaeoglobus profundus	1	42	1819	4.0	82	0.42	0.43	0.650*	(0.188/-0.206)
Ferroglobus placidus	1	43	2480	2.8	85	0.44	0.48	0.716*	(0.239/-0.258)
Halobacteriales									
Haloarcula marismortui	3	49	3131	10.0	37	0.62	0.75	1.032*	(0.338/-0.366)
Halobacterium salinarum	1	47	2136	4.0	37	0.68	0.87	0.959*	(0.729/-0.637)
Halorubrum lacusprofundi	3	51	3184	11.1	31	0.66	0.86	0.536*	(0.432/-0.434)
Halomicrobium mukohataei	2	41	3173	- -	45	0.66	0.85	1.168*	(0.581/-0.591)
Halorhabdus utahensis	1	40	2998	- -	50	0.63	0.79	1.430*	(0.429/-0.439)
Haloterrigena turkmenica	3	42	3739	1.5	51	0.67	0.87	0.925*	(0.365/-0.407)
Haloquadratum walsbyi	2	45	2823	24.0	37	0.48	0.42	0.080	(0.348/-0.233)
Natrialba magadii	2	40	3559	12.0	37	0.62	0.76	1.081*	(0.349/-0.403)
Natronomonas pharaonis	1	46	2675	2.1	44	0.63	0.77	1.254*	(0.287/-0.263)
Methanobacteriales									
Methanobacterium thermoautotrophicus	2	39	1855	1.8	65	0.50	0.54	0.848*	(0.295/-0.302)
Methanobrevibacter ruminantium	2	50	2217	16.8	37	0.36	0.30	0.885*	(0.439/-0.431)
Methanobrevibacter smithii	2	36	1795	4.4	37	0.31	0.18	0.899*	(0.357/-0.408)
Methanosphaera stadtmanae	4	42	1534	5.3	37	0.28	0.11	0.976*	(0.426/-0.398)
Methanopyrales									
Methanopyrus kandleri	1	34	1691	0.8	98	0.61	0.74	0.957*	(0.490/-0.447)
Methanococcales									
Methanocaldococcus jannaschii	2	36	1682	0.5	83	0.31	0.22	1.138*	(0.306/-0.205)
Methanocaldococcus fervens	2	36	1546	- -	- -	0.32	0.23	1.178*	(0.254/-0.258)
Methanocaldococcus vulcanius	2	35	1727	0.8	80	0.33	0.24	0.489*	(0.238/-0.285)
Methanococcus aeolicus	2	36	1490	2.0	46	0.30	0.20	0.923*	(0.258/-0.293)
Methanococcus maripaludis S2	3	37	1772	2.3	37	0.33	0.23	1.805*	(0.299/-0.298)
Methanococcus vannielii	4	37	1678	5.8	38	0.31	0.20	1.461*	(0.308/-0.271)
Methanosarcinales									
Methanococcoides burtonii	4	50	2273	20.0	22	0.41	0.37	0.910*	(0.273/-0.309)
Methanosaeta thermophila PT	2	47	1696	- -	65	0.54	0.64	0.665*	(0.248/-0.336)
Methanosarcina acetivorans	3	60	4524	24.0	38	0.43	0.44	0.852*	(0.298/-0.315)
Methanosarcina barkeri fusaro	3	63	3607	6.9	34	0.39	0.37	0.671*	(0.263/-0.254)
Methanosarcina mazei	3	57	3371	10.0	34	0.42	0.41	0.892*	(0.233/-0.252)
Rice Cluster I MRE50	3	53	3184	- -	37	0.55	0.70	1.058*	(0.424/-0.416)
Methanomicrobiales									
Methanocorpusculum labreanum	3	53	1739	13.0	37	0.50	0.56	0.884*	(0.350/-0.396)
Methanoculleus marisnigri JR1	1	49	2489	- -	40	0.62	0.80	0.470*	(0.465/-0.630)
Methanoregula boonei	1	49	2450	48.0	37	0.55	0.64	0.690*	(0.284/-0.341)
Methanospirillum hungatei	4	51	3139	14.0	37	0.45	0.43	0.628*	(0.274/-0.294)
Methanocella paludicola	2	39	3004	101.0	37	0.57	0.74	1.241*	(0.471/-0.532)
Methanosphaerula palustris	3	55	2655	28.0	30	0.55	0.67	0.340	(0.417/-0.486)

Species	Gene Numbers[b]			GenT[c]	Temp[d]	GC content[e]		S[f]	Random[g]
	rRNA	tRNA	ORFs			(i)	(ii)		
Euryarchaeaota									
Thermococcales									
Pyrococcus abyssi	1	46	1788	0.6	96	0.45	0.49	0.817*	(0.276/-0.279)
Pyrococcus furiosus	1	46	2208	0.6	99	0.41	0.37	0.620*	(0.242/-0.232)
Pyrococcus horikoshii	1	46	2061	0.5	95	0.42	0.41	0.474*	(0.228/-0.218)
Thermococcus gammatolerans	1	44	2157	1.6	88	0.54	0.68	0.679*	(0.346/-0.398)
Thermococcus kodakarensis	1	46	2306	0.7	85	0.52	0.64	1.541*	(0.400/-0.392)
Thermococcus onnurineus	1	46	1976	- -	80	0.51	0.63	1.022*	(0.324/-0.316)
Thermococcus sibiricus	1	44	2037	0.7	78	0.40	0.36	0.079	(0.246/-0.213)
Thermoplasmatales									
Picrophilus torridus	1	47	1535	6.0	60	0.36	0.34	0.607*	(0.221/-0.238)
Thermoplasma acidophilum	1	45	1509	2.5	58	0.46	0.52	0.451*	(0.286/-0.298)
Thermoplasma volcanium	1	45	1499	2.5	60	0.40	0.39	0.120	(0.205/-0.215)
Aciduliprofundum boonei	1	31	1544	- -	- -	0.39	0.35	0.142*	(0.254/-0.267)
Crenarchaeota									
Desulfurococcales									
Aeropyrum pernix	1	51	2694	3.3	90	0.56	0.66	0.356*	(0.321/-0.324)
Hyperthermus butylicus	1	47	1602	2.0	99	0.54	0.57	0.649*	(0.257/-0.244)
Ignicoccus hospitalis	1	47	1434	1.0	90	0.57	0.75	1.039*	(0.330/-0.275)
Staphylothermus marinus	1	46	1570	- -	88	0.36	0.26	0.223*	(0.205/-0.205)
Desulfurococcus kamchatkensis	1	47	1471	- -	82	0.45	0.47	0.006	(0.197/-0.180)
Thermoproteales									
Caldivirga maquilingensis	1	42	1963	8.0	83	0.43	0.43	-0.013	(0.236/-0.199)
Pyrobaculum aerophilum	1	46	2587	3.0	98	0.51	0.58	0.077	(0.204/-0.212)
Pyrobaculum arsenaticum	1	46	2298	3.7	95	0.55	0.66	-0.024	(0.265/-0.273)
Pyrobaculum calidifontis	1	46	2149	5.5	93	0.57	0.70	-0.077	(0.293/-0.279)
Pyrobaculum islandicum	1	46	1978	2.7	98	0.50	0.51	-0.248	(0.362/-0.308)
Thermoproteus neutrophilus	1	46	1966	- -	85	0.60	0.76	0.092	(0.373/-0.433)
Thermofilum pendens	1	45	1824	- -	88	0.58	0.72	-0.275	(0.446/-0.398)
Sulfolobales									
Metallosphaera sedula	1	46	2256	4.5	65	0.46	0.51	-0.133	(0.229/-0.223)
Sulfolobus acidocaldarius	1	49	2292	5.7	70	0.37	0.31	-0.184	(0.228/-0.201)
Sulfolobus islandicus	1	41	2738	- -	75	0.35	0.28	-0.058	(0.199/-0.205)
Sulfolobus solfataricus	1	46	2977	6.0	78	0.36	0.31	-0.258	(0.234/-0.232)
Sulfolobus tokodaii	1	46	2826	6.0	75	0.33	0.23	-0.105	(0.258/-0.248)
Thaumarchaeota									
Nitrosopumilus maritimus	1	44	1795	30.0	28	0.34	0.20	0.745*	(0.315/-0.249)
Cenarchaeum symbiosum A	1	46	2017	- -	10	0.58	0.68	0.532*	(0.301/-0.441)
Korarchaeota									
Korarchaeum cryptofilum	1	46	1602	- -	83	0.49	0.57	0.181	(0.259/-0.295)
Nanoarchaeota									
Nanoarchaeum equitans	1	44	552	0.75	90	0.32	0.23	0.185*	(0.164/-0.172)

Table 5-1 The strength of selected codon usage bias in Archaea

[a]Species; [b]total numbers of ribosomal RNA operons, tRNA genes and predicted open reading frames per genome; [c]minimal doubling time in hours; [d]optimal growth temperature; [e]G+C content across (i) all positions (ii) synonymously variable third codon positions ; [f]the strength of selected codon usage bias; [g]95% null range of values for *S*. Dashes indicate absent values

Values of *S* are not randomly distributed among archaeal families (Table 5-1), and may indicate that related species of Archaea share patterns of selected codon usage bias simply due to shared ancestry. To examine patterns of phylogenetic non-independence, values of *S* are considered in their phylogenetic context (Figure 5-1). At the root of the tree, members of the Archaea are estimated to have shared common ancestry more than four billion years ago (BATTISTUZZI *et al.* 2004) (but note that there is large uncertaintly in this estimate), and even the most closely related species in this study, *Pyrococcus abyssi* and *Pyrococcus horikoshii*, share less than 80% of DNA sequence identity across 20 highly expressed genes, meaning that there is likely to have been substantial time for patterns of codon usage to diverge among lineages. The phylogeny of the Archaea was rooted at the midpoint, and is divided into two major phyla; the Euryarchaeota and Crenarchaeota (either side of root in Figure 5-1). Strikingly, all members of the Crenarchaeota exhibit low *S* values (Table 5-1), with non-significant values observed among members of two of the major crenarchaeote families, the Thermoproteoles and Sulfolobales. Given the ancient common ancestry of this clade, estimated to be around 3.5 billion years ago (BATTISTUZZI *et al.* 2004), this might imply that patterns of selected codon usage bias evolve very slowly. Yet in other clades, the strength of selected codon usage bias can vary considerably over much shorter phylogenetic distances. For instance the *S* value for *Thermococcus kodakarensis* ($S = 1.541$) is more than double the value of its sister taxon *Thermococcus gammatolerans* ($S = 0.679$), and *S* values also vary substantially for the two most closely related species in these analyses *P. abyssi* ($S = 0.817$) and *P. horikoshii*, ($S = 0.474$).

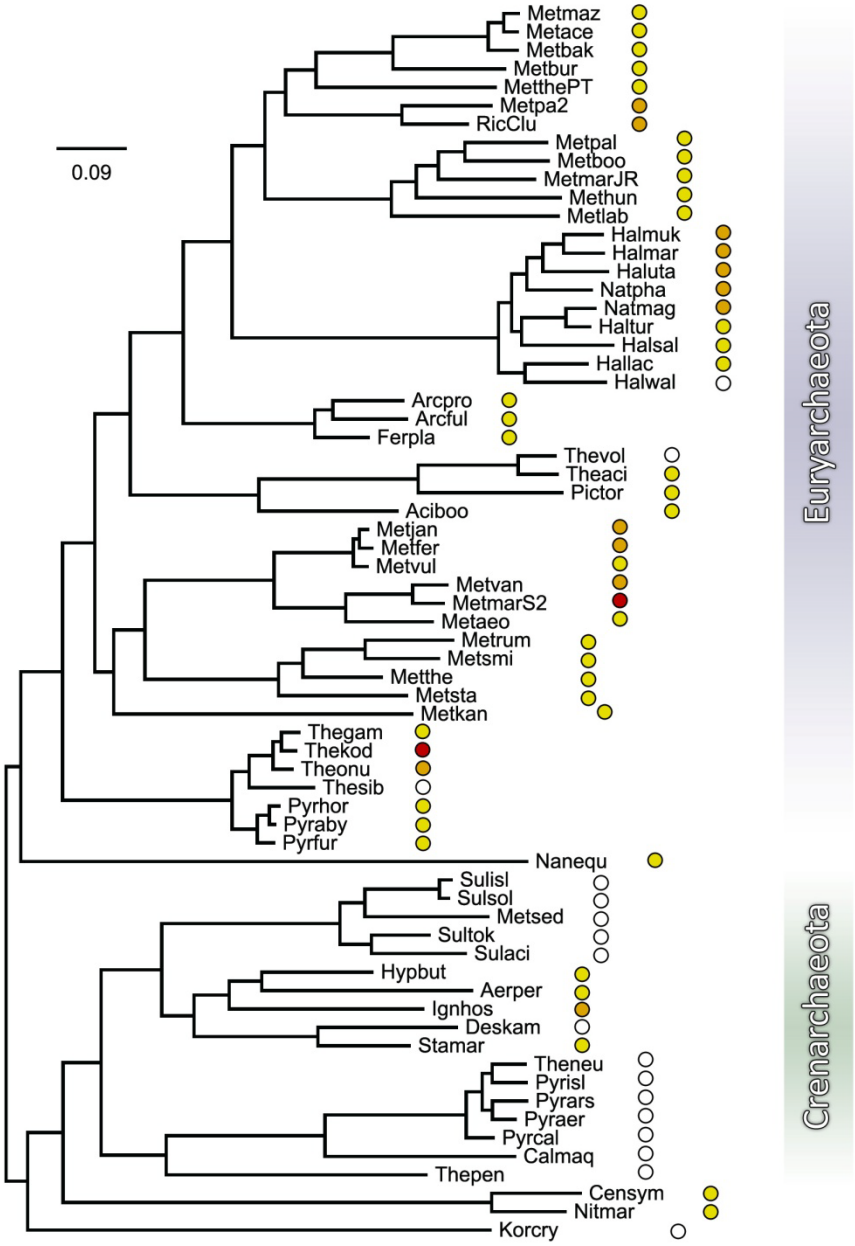


Figure 5-1 Phylogeny of Archaea

Midpoint rooted ML phylogeny for 67 species of Archaea constructed from the protein sequences of fourteen genes (see Methods) under the WAG model of protein evolution (WHELAN and GOLDMAN 2001) with gamma distributed rates among sites. The scale bar indicates the number of amino acid substitutions per site. The root is positioned at the midpoint. Coloured circles indicate species' *S* values: red $S > 1.5$; orange $S > 1$; yellow where *S* is significantly >0, and white where values are not. Species' codes indicated in AppendixA.

The strength of selected codon usage bias is not expected to be negative for any species. In bacterial genomes negative *S* values appear to reflect either (i) the position of highly expressed genes upon the C-poor leading strand in genomes with strong asymmetrical mutational biases, or (ii) abnormal base composition of highly expressed genes within 'genomic islands' (SHARP *et al.* 2005). Several negative values were observed among Archaea, and some appear to have a different origin to those observed among Bacteria. Unlike bacterial genomes, there can be substantial regional variation in GC3s within archaeal genomes (Chapter 4). This is the case for the genomes of *T. pendens* and *Sulfolobus acidocaldarius* where highly expressed genes are located within the A+T rich regions of the chromosome and coincide with the location of the replication origin-associated *cdc6* and *cdv* genes (LINDÅS *et al.* 2008; LUNDGREN *et al.* 2004). For instance, although *S. acidocaldarius* exhibits an average genome-wide GC3s of 0.31, patterns of GC3s vary across the chromosome with three-phase periodicity, and are lowest at the origins of replication. Most (14) of the 20 highly expressed genes analysed are located close to the second origin of replication (located around gene position 600 across the sequenced chromosome) and are thus relatively A+T rich (mean GC3s = 0.26). Yet three of the other highly expressed genes located in a different genomic region have higher GC3s values (mean = 0.32). Heterogeneity in GC3s among highly expressed such as this cannot be explained by translation selection for optimal codons because the same optimal codons that best correspond to the most abundant tRNA species are expected to be selected across all highly expressed genes. In this case, the A+T richness of most highly expressed genes results in a lower frequency of C-ending optimal codons, and so a lower (and in this case negative) estimate of *S* (Equation 2-12).

5.3.2 Correlations of *S* with genome characteristics in Archaea

Four correlations of genome characteristics are observed among bacterial genomes. Interspecific numbers of rRNA operons and tRNA genes are highly correlated (KANAYA *et al.* 1999; SHARP *et al.* 2005), and the strength of selected codon usage bias is correlated with the numbers of ribosomal RNA operons, the numbers of transfer RNA genes and generation time (SHARP *et al.* 2005; SHARP *et al.* 2010; VIEIRA-SILVA and ROCHA 2010), as expected if these traits coevolve for the efficiency of growth (BERG and KURLAND 1996; EHRENBERG and KURLAND 1984). If the genomes of Archaea are subject to the same global selective pressures as bacterial genomes, then the same correlations might be expected to be observed among Archaea. Trends observed among Bacteria were superimposed with values for the Archaea

(Figure 5-2) to investigate if Archaea exhibit similar patterns. Consistent with this view, values for Archaea generally fall within the distribution of values for Bacteria. Figure 2A illustrates that Archaea have a restricted range of rRNA operons (1-4) and tRNA genes (31-63) compared with Bacteria but the relationship of the two traits across both domains is similar. The range of minimal generation times is more restricted across Archaea by comparison with Bacteria (Figure 2B), but the average values are similar for both domains (10.8 and 9.4 days respectively). Consistent with the lower numbers of rRNA operons and tRNA genes, the strength of selected codon usage bias is reduced in range in Archaea by comparison with bacterial species, although there appears to be a slight excess of high *S* values in species with low numbers of rRNA operons and tRNA genes (Figure 2C and D). Across Archaea, independent contrasts for the numbers of tRNA genes and rRNA operons are positively correlated ($r = 0.32$, $p < 0.01$), although independent contrasts for values of *S* are not significantly correlated with either numbers of rRNA operons ($r = 0.02$, $p = 0.58$), or tRNA genes ($r = 0.12$, $p = 0.13$), and might reflect a lack of statistical power associated with the low range of values and numbers of contrasts. There was no significant correlation between independent contrasts of the strength of selected codon usage bias and generation time across these Archaea ($r = 0.11$, $p = 0.35$).

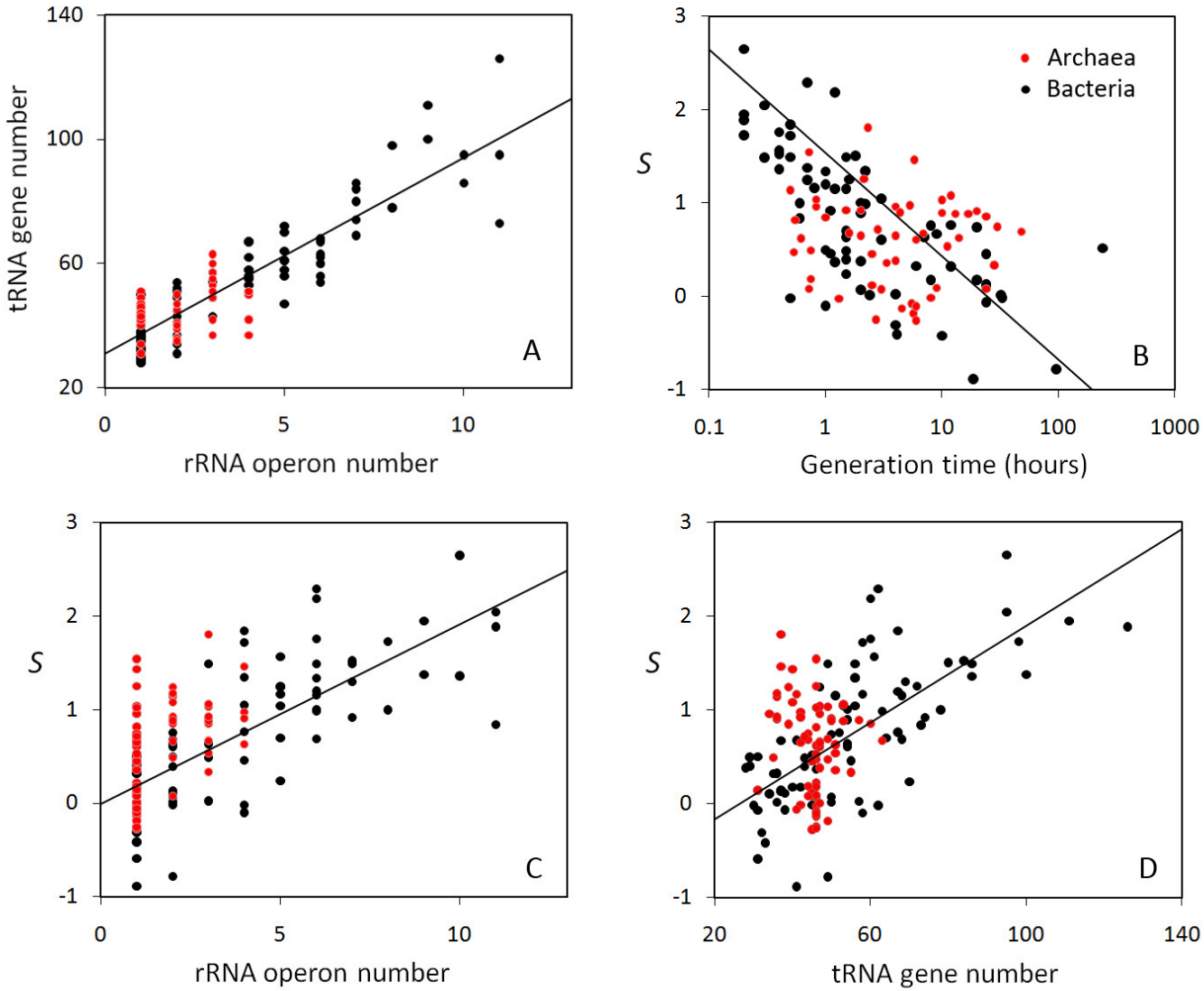

140
100
60
20
tRNA gene number
0
5
10
rRNA operon number
A
3
2
1
0
-1
S
Archaea
Bacteria
0.1
1
10
100
1000
Generation time (hours)
B
C
D

Figure 5-2 Correlations of genome characteristics in Archaea and Bacteria

Black points indicate Bacteria. Red points indicate Archaea. A: Correlation of the total numbers of tRNA genes and rRNA operons per genome. B: Correlation of the strength of selected codon bias with minimal generation time among Bacteria but not Archaea. C&D: Strength of selected codon usage bias as a function of the numbers of (C) rRNA operons and (D) tRNA genes. Regression lines for bacterial trends are shown.

A linear model was used to identify factors from rRNA operon number, tRNA gene number and minimal generation time that explained variation in independent contrasts of *S*. However, unlike the situation observed for bacterial genomes (SHARP *et al.* 2005; SHARP *et al.* 2010; VIEIRA-SILVA and ROCHA 2010), none of these factors when considered in isolation or combination were found to explain significant variance in contrasts of *S*. Unlike many of the bacterial species that have been subject to previous investigations, more than half of these Archaea grow optimally at temperatures greater than 40˚C. Since growth temperature is expected to have a kinetic impact upon growth rate (DETHLEFSEN and SCHMIDT 2005; FAREWELL and NEIDHARDT 1998), it was considered as the fourth factor in a linear model to identify factors explaining variation in independent contrasts of *S*. These analyses revealed that only two factors, minimal generation time and optimal growth temperature contribute significantly (both $p < 0.01$), each accounting for a similar proportion of variance in *S* (10.2% and 8.1% respectively). To investigate these factors further, species were binned into two arbitrary categories according to growth temperature of above and below 40°C (Figure 5-3). In each category, values (both raw and independent contrasts) were negatively correlated with generation time, with species in the high growth temperature category associated with a systematic reduction in the strength of selected codon usage bias. This result remains qualitatively unchanged when alternative temperature thresholds are considered. The effect of temperature is sufficient in magnitude to obscure any relationship of *S* with generation time alone (Figure 2B), and whilst the distributions of growth rates among Archaea and Bacteria are relatively similar, their distributions of growth temperatures are not. More than half of the Archaea in these analyses grow optimally at temperatures of 40°C but this fraction seems to be much smaller across the bacterial domain where only 5/80 species are known to grow optimally at such high temperatures. There is no significant difference in values of *S* between Archaea and Bacteria when only those species growing optimally below temperatures of 40°C (or similar thresholds) are considered (t-test, $p = 0.23$).

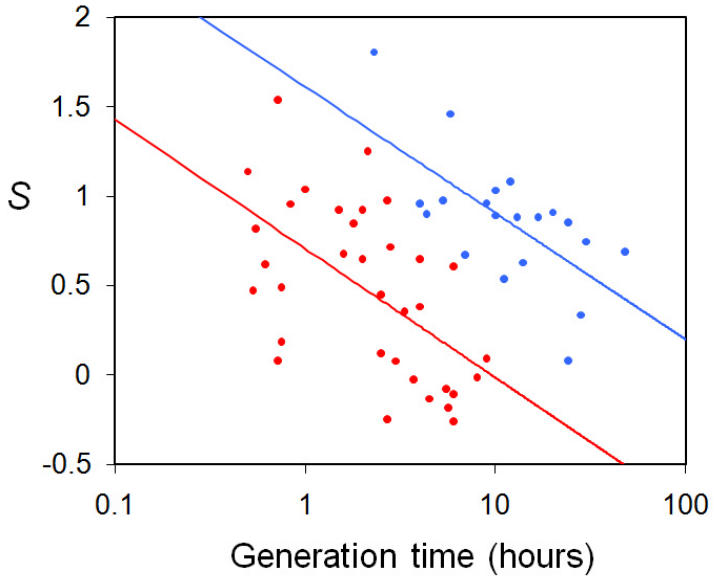


Figure 5-3 Variation in the strength of selected codon usage bias with generation time and optimal growth temperature in Archaea

The strength of selected codon usage bias (*S*) as a function of generation time and optimal growth temperature for 52 species of Archaea. Red points indicate optimal growth temperatures > 40°C ($n = 32$). Blue points indicate optimal growth temperatures < 40°C ($n = 20$). Lines indicate regression of *S* on generation time for each dataset ($r = 0.43$, $p < 0.01$; $r = 0.31$, $p < 0.01$ of independent contrasts for each temperature category respectively).

5.3.3 Impact of growth temperature upon genome characteristics in Archaea

Here, optimal growth temperature has been shown to impact upon patterns of selected codon usage bias in Archaea, and so might be expected to impact upon other aspects of genome evolution. The original plots of the relationship of rRNA operon numbers, tRNA gene numbers and *S* (Figure 5-2A, C and D) were not found to be illuminated by categorical binning according to optimal growth temperature since most thermophiles contain a single rRNA operon and a minimal tRNA gene set. Variation in phylogenetic independent contrasts of optimal growth temperature was examined with respect to contrasts for generation time, numbers of rRNA operons, numbers of tRNA genes and genome size. All four traits were found to be negatively correlated with growth temperature (Figures 5-4 and 5-5; $r = 0.71$, $p > 0.001$; $r = 0.60$, $p > 0.001$; $r = 0.26$, $p = 0.03$; $r = 0.37$, $p < 0.01$ respectively). Futhermore, contrasts of generation time were found to be positively correlated with those of chromosome size ($r = 0.38$, $p < 0.01$; Figure 5-5C).

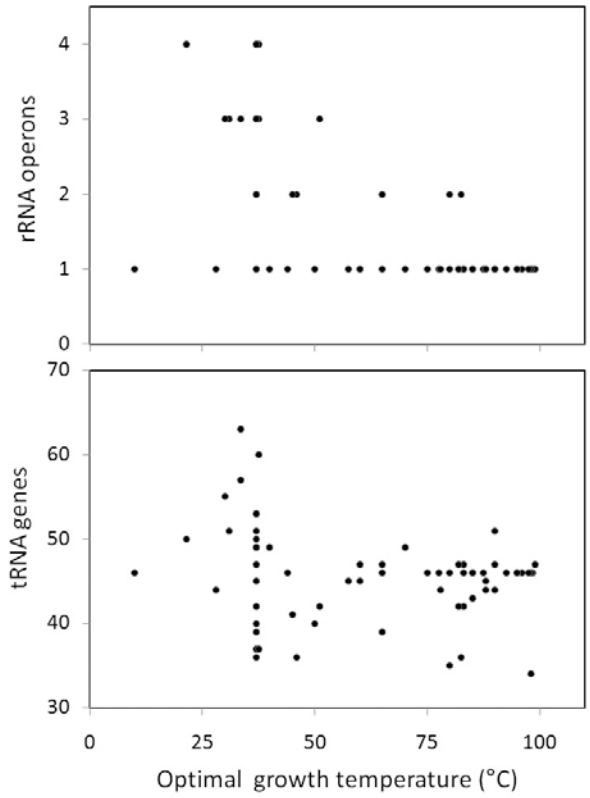


Figure 5-4 Impact of optimal growth temperature upon rRNA operons and tRNA genes in Archaea

Numbers of rRNA operons and tRNA genes as a function of optimal growth temperature among 64 species of Archaea.

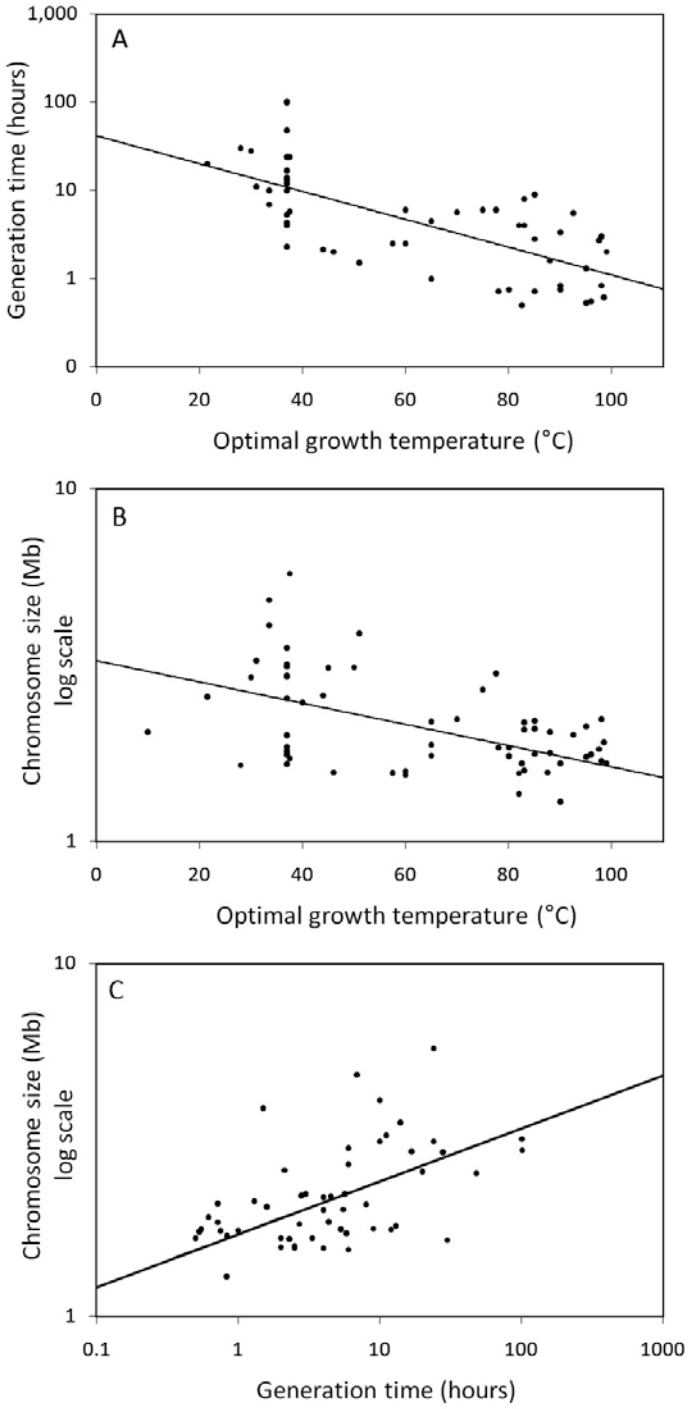


Figure 5-5 Correlations of generation time, optimal growth temperature and genome size and among Archaea

(A) Correlation of optimal growth temperature with generation time among 52 Archaea. (B) First chromosome size as a function optimal growth temperature for 65 species of Archaea. (C) First chromosome size as a function of generation time for 52 species of Archaea.

5.4 Discussion

5.4.1 Comparison with Bacteria

The strength of selected codon usage bias (*S*) among Archaea is reduced in range (-0.28 – 1.81) by comparison with bacterial species (-0.88 – 2.65), consistent with the reduced range of ribosomal RNA operons and tRNA genes across Archaea, and as expected if there is a common strategy for the efficiency of competitive growth among microorganisms (BERG and KURLAND 1996; EHRENBERG and KURLAND 1984). A similar proportion of bacterial (30%) and archaeal (28%) genomes exhibit non-significant *S* values, thus the lower range of values across Archaea appears to reflect a deficit of high values typically associated with a generalist lifestyle among Bacteria. Consistent with this view, 4/67 Archaea are classified as inhabiting multiple environments compared with 28/80 Bacteria according to NCBI (Appendix C). Among Bacteria, low and non-significant *S* values are typically associated with the lifestyles of endosymbionts where exponential growth does not contribute substantially to life history. No symbioses of Archaea have been reported with the exception of *Nanoarchaeum equitans* – an obligate symbiont of *Ignicoccus hospitalis* (HUBER *et al.* 2002), which exhibits weak selected codon usage bias ($S = 0.185$) thus conforming to the bacterial trend. However, host-dependency does not appear to be the prevailing factor explaining the reduced range of *S* values among these Archaea since growth rates of endosymbionts are typically slow, and yet here there is no relationship of *S* and growth rate when these variables are considered in isolation (Figure 5-2B). Rather, variation in *S* was found to be explained by the combined effects of growth temperature and growth rate. Since the distribution of growth rates across Archaea and Bacteria are similar but the distributions of growth temperatures are not, it seems that the generally lower values of *S* among Archaea reflects the large fraction of Archaea with high optimal growth temperatures. Consistent with this, values of *S* were not significantly different between the domains when comparisons were restricted to species with low growth temperatures.

5.4.2 Kinetic impact of growth temperature upon growth rate in Archaea

The observed systematic reduction in values of *S* at high temperatures is not entirely unexpected, and might be explained by a number of processes. First, the observation is expected if there is an inverse relationship with effective population size (N_e) and optimal

growth temperature since S is expected to reflect twice the product of the effective population size and the selection coefficient upon codon usage bias ($2N_eS$). However the impact of effective population size is expected to reduce the efficacy of all forms selection, and yet selection upon protein sequences appears to be increased for thermophiles where non-synonymous to synonymous (d_N/d_S) ratios across orthologous gene sequences are consistently lower (FRIEDMAN *et al.* 2004), and so this observation appears to be inconsistent with an effect of N_e. Second, values of S might be reduced at high temperatures if there is a reduction in competitive growth in such specialised environments. In this case, species growing optimally at higher temperatures are expected to exhibit slower growth rates, and yet here the opposite was found to be the case ($r = 0.71$, $p < 0.001$; Figure 5), and so it seems that a temperature-associated reduction in growth competition is incapable of explaining the reduced values of S. Finally, values of S might be lower due to the kinetic impact of temperature upon the rate of translation elongation. Kinetic theory predicts that the rate of any chemical reaction that requires activation energy (such as translation elongation) is expected to vary with temperature following a predictable relation:

$$v = K e^{\Delta E/RT}$$

Where v is the rate of reaction, K is a rate specific constant, ΔE is the activation energy, R is the universal gas constant and T is the temperature. Thus growth temperature is expected to be directly proportional to the log of elongation rate, and this relation has been confirmed *in vivo* for the bacterium *Escherichia coli* (FAREWELL and NEIDHARDT 1998). There is some indication that elongation rates are to some extent growth limiting, as they are linearly related to growth rate across viable growth temperatures (FAREWELL and NEIDHARDT 1998), thus leading to the expectation that the log of generation time is proportional to growth temperature. This prediction has led some authors to correct the growth rate by growth temperature in comparative analyses (DETHLEFSEN and SCHMIDT 2005). Here, this predication has been confirmed across 54 species of Archaea (Figure 5-5A), consistent with the kinetic modulation of growth rate among species.

5.4.3 Implications for the nature of translational selection

The impact of optimal growth temperature upon the strength of selected codon usage bias may shed light upon the debate surrounding the selective benefit of optimal codons. A variety of biochemical changes occur at high temperatures which may have a substantial

impact upon DNA sequence evolution. Rates of protein degradation (LIN and ZABIN 1972; ST JOHN and GOLDBERG 1978) and intracellular concentrations of molecular chaperone proteins (GOFF *et al.* 1984) are elevated. Chaperone abundances are higher in thermophilic species such as *Pyrodictium occultum,* where GroEL (a chaperone which aids correct protein folding), is present at eight-fold higher concentrations (PHIPPS *et al.* 1993) by comparison with species with moderate growth temperatures such as *E. coli* (VAN BOGELEN *et al.* 1992) or *Buchnera aphidicola* (BAUMANN *et al.* 1996). These observations are indicative of a greater burden of misfolded proteins at high temperatures. Any deleterious effects of protein misfolding may be minimised by selection for particular amino acid sequences which are robust to further substitution (WILKE and DRUMMOND 2006). As expected if this form of selection is greatest in species with high optimal growth temperatures, orthologous protein sequences appear to be under more intense purifying selection among thermophiles (FRIEDMAN *et al.* 2004). Collectively these observations indicate that it is more important to obtain the correct amino acid sequence at high temperatures, meaning that errors in translation are more costly, and so selection for optimal codons to avoid translation errors is expected to be greater. Yet, here across 54 species of Archaea spanning a temperature range from 10 – 99°C, the strength of selected codon usage bias was found to be reduced rather than elevated among thermophiles (Figure 5-3). It therefore seems unlikely that the primary target of selection among Archaea is for the accuracy of translation.

Alternatively, the benefit of selected codon usage bias may be for the efficiency of translation, if optimal codons are translated more quickly than their synonyms and increase the rate of translation elongation. Under this hypothesis, each ribosome spends less time translating any particular transcript, and is thus more efficient, able to translate a larger number of mRNAs in a given time, and so the overall cellular rate of translation is increased (EHRENBERG and KURLAND 1984). Then, provided that the rate of translation is to some extent growth rate limiting, an increase in translation rate confers a growth advantage.

The kinetic effect of growth temperature upon the rate of translation elongation is likely to impact upon this process via two mechanisms. First, growth temperature increases the rate of translation elongation, necessarily decreasing the time spent translating each codon, and so the time saved in translating an optimal codon over its synonym. At higher temperatures, the time saved in the usage of optimal codons at any particular growth rate is a smaller fraction of generation time, and thus of less selective benefit. Second, since the efficiency

saving in optimal codon usage is effectuated via efficient utilisation of ribosomes; its benefit to any particular species is expected to depend upon the fraction of resources invested in ribosomes rather than other cellular machinaries. For some species this can be rather high; for instance, within rapidly growing *E. coli* cells, two thirds of the protein mass is accounted for by ribosomes (PEDERSEN *et al.* 1978). For other species, such as those living at high growth temperatures, this fraction is expected to be lower. At high temperatures, ribosomes are intrinsically more efficient; rates of elongation proceed rapidly in the absence of any increase in ribosome or ternary complex concentration (FAREWELL and NEIDHARDT 1998). Therefore any particular growth rate may be attained with less ribosomal investment, thus reducing the selective benefit in optimal codon usage. Consistent with this, the vast majority (80%) of the cellular protein fraction for the thermophile *Pyrodictium occultum* is not accounted for by ribosomes (PHIPPS *et al.* 1993). Furthermore, the negative correlation of the numbers of rRNA operons with growth temperature (Figure 5-4) suggests that there is indeed less investment in ribosomes among Archaea growing at high temperatures. As expected if reduced ribosomal investment decreases the selective benefit in optimal codon usage for global cellular efficiency, the strength of selected codon usage bias (*S*) is systematically lower at higher growth temperatures for any particular growth rate (Figure 5-3). Thus it seems based upon these observations, that it is selection for translation efficiency, rather than accuracy, that primarily shapes selected codon usage bias among Archaea.

From the above lines of enquiry, one might be left wondering why, under the efficiency model, Archaea growing at high temperatures are not under ever-more selection to attain ever-more rapid translation rates, in order to attain ever-faster growth rates via rRNA duplication and/or selected codon usage bias. However, this seems to be a likely consequence of the kinetic impact of growth temperature upon different cellular processes. Whilst rates of translation elongation are expected to increase with increasing growth temperature, the same is not true for other cellular processes such as DNA synthesis. Nucleotide substrates are incorporated as activated precursors and so their ΔE is zero (see equation 5.4.2), meaning that increasing growth temperature is expected to have no impact upon the rate of DNA synthesis (XIA 1995). Thus, as growth temperature increases, the rate of translation elongation is expected to become less rate limiting, relinquishing control to other processes such as DNA replication. Indeed, beyond 44°C, the rate of translation elongation is no longer growth rate limiting for *E. coli* (FAREWELL and NEIDHARDT 1998). Thus under an efficiency model, selection for characteristics such as codon usage bias, rRNA

operons and tRNA genes, which can increase the rate of translation, are likely to be of diminishing benefit as translation has less impact upon growth rate. This is in support of other predictions made by the efficiency model and consistent with trends observed among Archaea. Furthermore, the expected increase in importance of DNA replication in limiting cell growth at high temperatures has the potential to explain why chromosome size decreases with growth temperature (Figure 5-5B), as selection for rapid growth leads to a reduction in chromosome size (Figure 5-5C).

5.4.4 Genome evolution at high temperatures

Whilst the notion of adaptive evolution of genome size remains a controversial issue, it seems difficult to explain reduced chromosome sizes at high temperatures among these Archaea as a consequence of neutral evolution. Among bacterial genomes, small genomes are considered to reflect the inability of natural section to counteract mutational pressure towards gene deletion (MIRA *et al.* 2001), although the relevance of these effects appears to be restricted to the genomes of endosymbionts (MORAN *et al.* 2009), of which there are few among these Archaea.

One possibility, which is consistent with the theory of translational efficiency (described above), is that selection favours gene deletions at high temperatures to minimise the time spent replicating a chromosome, to confer a growth advantage. Consistent with this theory, among Archaea, small chromosome sizes are associated with low numbers of rRNA operons, high growth temperatures and shorter generation times (Figure 5-5B&C).

Alternatively, these same observations might be explained if selection acts to shape mutation rates (DRAKE 1991). Mutation rates among thermophilic species are around an order of magnitude lower than species growing optimally at moderate temperatures (DRAKE 2009). If mutation rate is selected then these observations could reflect an increased cost of mutation with increasing growth temperature. Any reduction in genome size is expected to lower the cost of mutation since it reduces the number of mutational targets. Thus at high temperatures where the cost of mutation appears to be greater, selection may favour smaller genome sizes to minimise this cost, consistent with observations (Figure 5-5B). It may be possible to better discriminate between these hypotheses (or others) when more comprehensive replication origin data becomes available for the Archaea, since the rate of

DNA synthesis is expected to vary with replication origin numbers but the numbers of mutational targets is not.

In conclusion, the strength of selected codon usage bias is reduced among Archaea by comparison with Bacteria, and appears to be the consequence of differences in growth temperature among the domains analysed here. Consistent with kinetic theory, growth temperature appears to have a marked impact upon genomic architecture, and it will be interesting to see if Bacteria living at high temperatures display similar trends.

6. TRENDS IN CODON USAGE BIAS AMONG ARCHAEAL GENOMES

6.1 Introduction

Patterns of codon usage can vary considerably among genomes. Across the bacterial domain, there is substantial heterogeneity in G+C content, which is greatest at third codon positions (MUTO and OSAWA 1987). A variety of multivariate analyses have established heterogeneity in G+C content as the primary source of variation in codon usage bias among bacterial genomes (CHEN *et al.* 2004; LOBRY and NECSULEA 2006; LYNN *et al.* 2002), presumably reflecting variation in patterns of mutation. These analyses also identified a second source of heterogeneity in codon usage bias, which was found to be associated with optimal growth temperature. For reasons that remain unclear, this source has often been interpreted as codon adaptation to high temperatures (LYNN *et al.* 2002; PUIGBO *et al.* 2008).

In Chapter 4 I determined and explored factors giving rise to variation in codon usage within the genomes of Archaea. Each of these factors: variation in (i) G+C content, (ii) asymmetric mutational biases, and (iii) the effectiveness of selection for translationally optimal codons, has the potential to impact upon heterogeneity in codon usage among the genomes of Archaea. Here, I employ multivariate analyses to establish the most important determinants of codon usage variation among Archaea. Four major trends were established. Consistent with patterns observed among the bacterial domain, the first and second were associated with variation in G+C content and growth temperature respectively. The nature of these trends was explored. Heterogeneity in G+C content was not related to growth temperature or aerobiosis and is most simply explained by the occurrence of species-specific mutational biases. The codon giving rise to most variation in the secondary trend, prevalent among thermophilic species, was found to be selectively disfavoured, and is thus most unlikely to reflect thermophilic adaptation. Furthermore, I provide some indication that the secondary trend is likely to reflect a context-dependent transcription-coupled mutational bias linked with growth temperature.

6.2 Methods

6.2.1 Identifying major trends in codon usage among the genomes of Archaea

The same datasets of synonymous codon usage (i) genome-wide and (ii) across highly expressed genes, together with data for optimal growth temperatures, for each of 67 species of Archaea were used as in Chapters 4 and 5. Optimal growth temperature data were available for 65/67 species. Whilst in Chapters 3 and 4, within-group correspondence analyses (WCA) were performed upon the summed codon usage for each gene within a genome to identify major trends among genes, here WCA is applied to the summed codon usage for each of 67 genomes, to identify the major trends among genomes. To explore trends in both the presence and absence of translational selection, WCA was performed upon two datasets for codon usage across (i) highly expressed genes and (ii) genome-wide, and was implemented in the ade4 package (CHARIF *et al.* 2005) of the R suite. The resulting WCA axes successively explain variation in codon usage among genomes. To explore the sources of variation underlying each axis, axes were correlated with a variety of genome characteristics and codon usage indices including: average base composition of synonymously variable third codon positions across all protein coding genes, the strength of selected codon usage bias (*S*), the average genome-wide enrichment of G over C and T over A across all protein coding genes (G-C and T-A skews) and optimal growth temperature.

6.2.2 Exploring trends

Phylogenetic independent contrasts (FELSENSTEIN 1985) were implemented in the ape package (PARADIS *et al.* 2004) of the R suite, to assess the impact of optimal growth temperature upon genomic G+C content. To assess the impact of aerobiosis upon G+C content, the aerobic status of each archaeal species was assigned according to NCBI classification (Appendix C). To investigate any potential selective effect upon the codon explaining most variation upon the second WCA axis, the strength of selected codon usage bias (*S*) was adapted for this codon (see Methods Chapter 2).

6.3 Results

6.3.1 Major trends in codon usage among Archaea

Multivariate analyses have revealed several trends describing patterns of codon usage among 67 species of Archaea. Three of these trends, and their relative importance, were consistently identified by independent analyses of genome-wide and highly expressed gene codon usage (Appendix D). Indeed genome coordinates upon primary, secondary and tertiary axes from these independent analyses were highly correlated ($r = 0.99$; $r = 0.96$; $r = 0.86$ respectively, all $p < 0.001$). The first axis accounted for 67% of variation in genome-wide codon usage and 58% of variation in highly expressed gene codon usage among species, and was highly correlated with genomic G+C content (both $r = 0.99$, $p < 0.001$). Other trends were relatively minor by comparison. The second axis explained 12% of variation in genome-wide and 13% of highly expressed gene codon usage among species and was correlated with optimal growth temperature ($r = 0.72$, $r = 0.73$ respectively, both $p < 0.001$). A plot of the first WCA axis against the second reveals clear separation on the basis of GC3s and optimal growth temperature; separation which appears to be clearer for the WCA analysis of highly expressed genes (Figure 6-1).

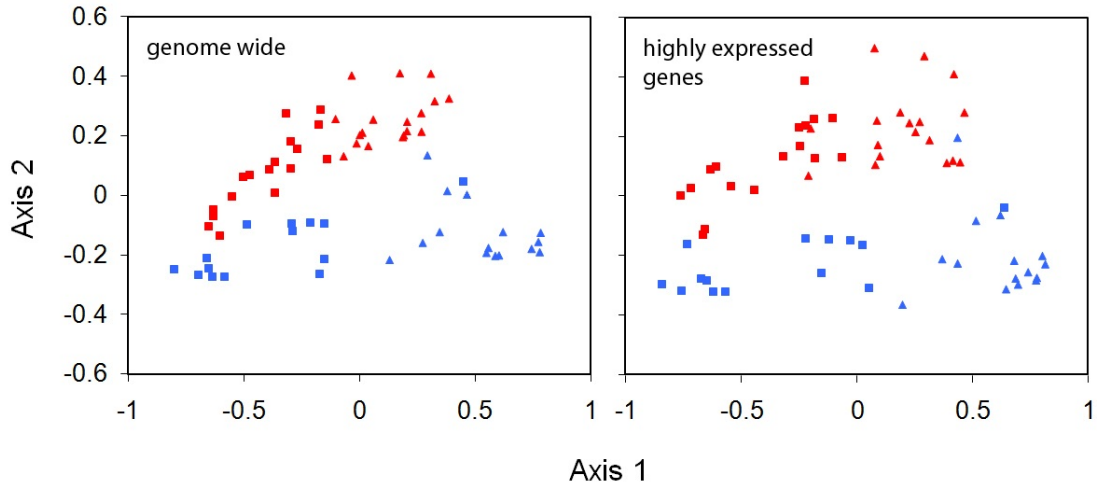


Figure 6-1 Positions of 67 species of Archaea upon primary and secondary within-group correspondence analyses axes

Analyses of summed codon usage of both genome-wide and highly expressed gene datasets. Triangles indicate species with greater than 50% G+C content at synonymously variable third codon positions (GC3s). Squares indicate species with less than 50% GC3s. Red indicates species living at temperatures of 50°C or warmer, and blue indicates cooler optimal growth temperatures.

Optimal growth temperature was positively correlated with average genome-wide G-C skew ($r = 0.71$, $p < 0.001$), and G-C skew correlated to a lesser extent with the secondary WCA axis ($r = 0.66$, $p < 0.01$). The variation on axis 2 appears to be primarily due a small subset of arginine and isoleucine codons, with the AGG arginine explaining most variation (Figure 6-2).

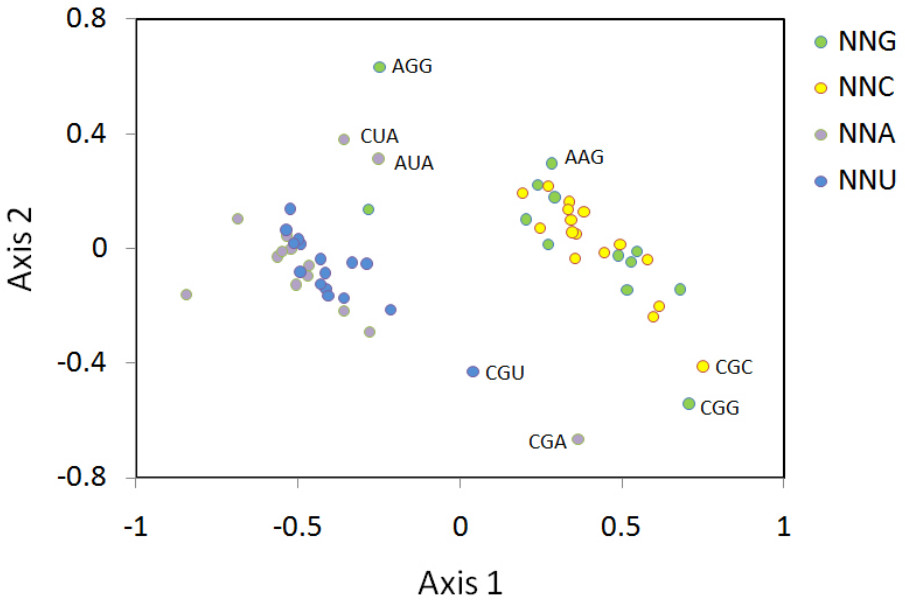


Figure 6-2 Positions of codons upon primary and secondary within-group correspondence analysis axes of 67 species of Archaea

Within-group correspondence analysis (WCA) performed upon the summed codon usage across all protein coding genes. Colours indicate the third codon position of each codon. Codons explaining the most variation upon axis 2 are labelled.

The third WCA axis explained relatively little variation in genome-wide (5%) and highly expressed gene (7%) codon usage among species and was correlated with average genome-wide T-A skew ($r = 0.60$, $p < 0.01$) and weakly with growth temperature ($r = 0.31$, $p < 0.05$). The fourth WCA axis was different in independent analyses of genome-wide and highly expressed gene codon usage; genome coordinates upon this axes in each analysis were not correlated ($r = 0.06$, $p = 0.42$). The fourth axis from WCA of genome-wide codon usage was not found to be correlated with any index and explained little variation among genes (3%).

Yet the fourth axis from WCA of codon usage across highly expressed genes was found to be correlated with the strength of selected codon usage bias (S) ($r = 0.56$, $p < 0.01$), with each of the four universally optimal codons (UUC, UAC, AAC and AUC) falling to one side of the distribution.

6.3.2 Exploring trends in codon usage among genomes of Archaea

Genome G+C content varies among Archaea from 28% in *Methanosphaera stadtmanae* to 68% in *Halobacterium salinarum* across species investigated here. As with bacterial genomes, Archaea possess little non-coding DNA, so with 1st and 2nd codon positions constrained by amino acid sequence requirements, most heterogeneity in G+C content among Archaea is at third codon positions. Consistent with this, genome-wide G+C content across synonymously variable third codon positions (GC3s) varies from 11% to 87% among Archaea. The effect is so marked that it can be observed among individual gene sequences e.g. *dnaG* (Figure 6-3).

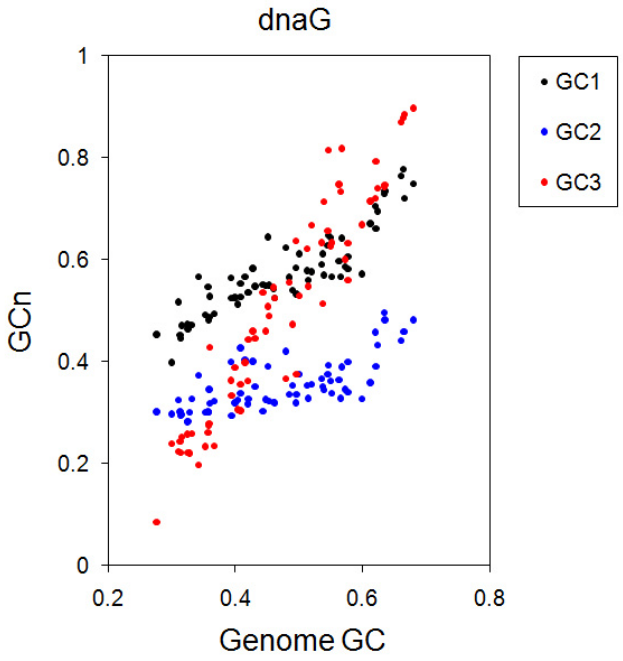


Figure 6-3 Correlation of *dnaG* G+C content at each codon position with genome-wide G+C content

Black points indicate G+C content at first codon positions. Blue points indicate G+C content at second codon positions. Red points indicate G+C content at third codon positions.

It has been suggested that there might be a selective benefit in the use of G+C bases at high temperatures, given their ability to form three (rather than two) hydrogen bonds (BERNARDI and BERNARDI 1986). Optimal growth temperatures vary substantially among Archaea from 10°C to 99°C in the species investigated here. No correlation of optimal growth temperature and genome-wide G+C content ($r = 0.05$, $p = 0.54$) or GC3s ($r = 0.02$, $p = 0.27$; Figure 6-4) was observed, even when considering the possibility that trends are obscured by the underlying phylogenetic relationships among archaeal species ($r = 0.04$, $p = 0.60$ and $r = 0.12$, $p = 0.18$ respectively).

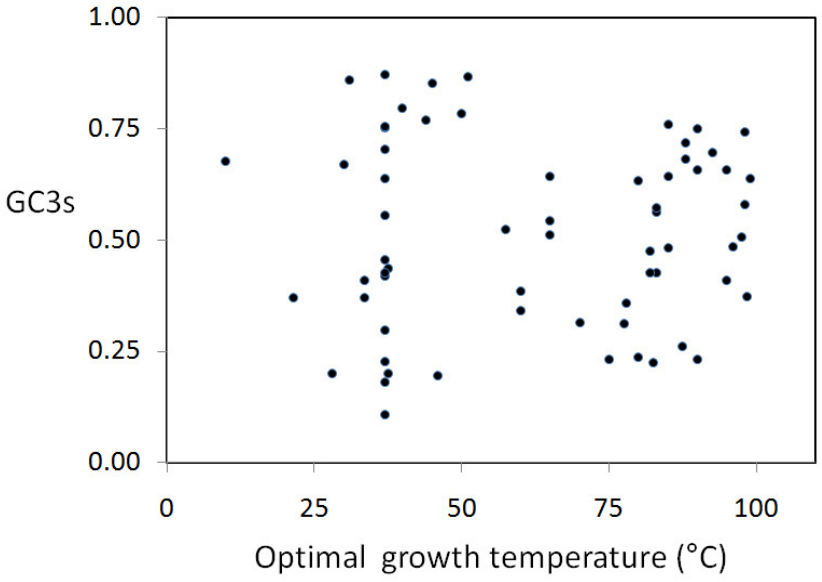


Figure 6-4 Relationship of G+C content at synonymously variable third codon positions with optimal growth temperature among 65 species of Archaea

Similarly, there has been speculation that natural selection has a role in shaping G+C content differently among bacterial aerobes and anaerobes (NAYA *et al.* 2002). There were 14 aerobes and 46 anaerobes among the Archaea examined here and their distributions of genome-wide G+C and GC3s values did not differ significantly (t-tests, $p = 0.45$ and $p = 0.61$ respectively, Figure 6-5).

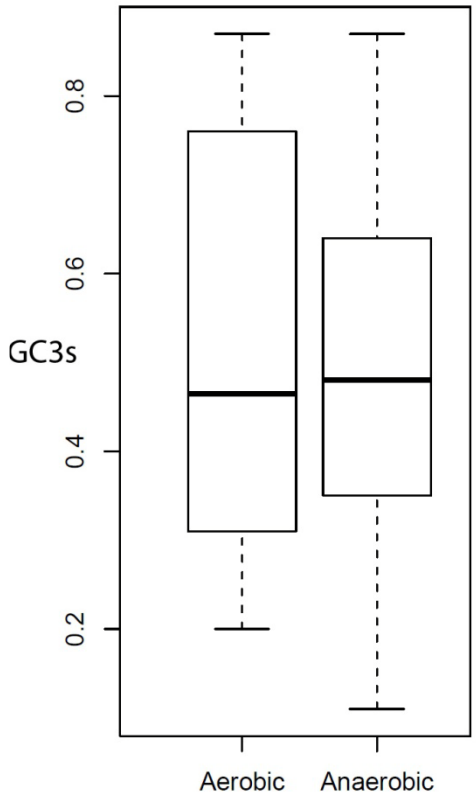


Figure 6-5 Relationship of G+C content at synonymously variable third codon positions (GC3s) with the status of archaeal aerobic metabolism

Dark horizontal lines indicate median values; boxed areas indicate the 50% range of values; error bars indicate the 95% range of values.

Multivariate analyses of bacterial genomes commonly identify a secondary trend associated with optimal growth temperature (LOBRY and CHESSEL 2003; LOBRY and NECSULEA 2006; LYNN *et al.* 2002), which has often been interpreted as reflecting an adaptation to high temperatures (LYNN *et al.* 2002; PUIGBO *et al.* 2008). Like the secondary WCA axis in these analyses, the trend is largely caused by variation in the arginine AGG codon which is enriched among thermophiles (LOBRY and NECSULEA 2006). To investigate any potential selective benefit of the AGG codon at high temperatures, its strength of selection was estimated (i.e. *S* for AGG denoted S_{AGG}). Values of S_{AGG} vary among species from -2.49 in *Haloarcula marismortui* to 0.75 in the uncultured Rice Cluster I MRE50, with an average value of -0.34, and nearly two thirds of species (24) exhibiting negative values. Whist S_{AGG} is only

based upon the use of a single codon, it accounts for a considerable proportion (35%) of variance in the secondary WCA axis, and is positively correlated with optimal growth temperature ($r = 0.33$, $p < 0.05$), although this relationship largely reflects a deficit of very negative S_{AGG} values at high temperatures (Figure 6-6).

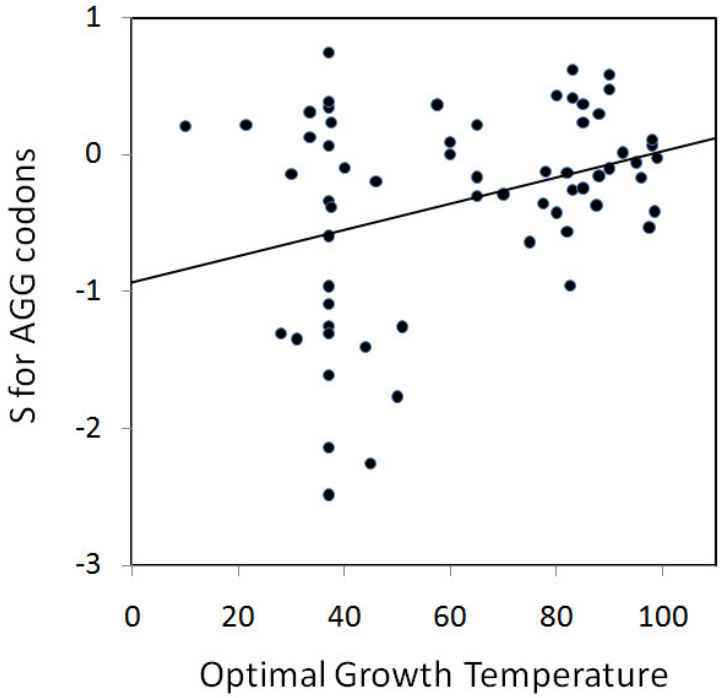


Figure 6-6 Correlation of the strength of selected arginine AGG usage bias with optimal growth temperature among 67 species of Archaea

6.4 Discussion

Four trends explain variation in codon usage among the genomes of Archaea. Previous studies of among-genome codon usage heterogeneity have mostly included species from the bacterial domain and have revealed only two major trends (CHEN *et al.* 2004; LOBRY and CHESSEL 2003; LOBRY and NECSULEA 2006; LYNN *et al.* 2002). Consistent with previous studies, this study has found that most variation in codon usage among species is explained by heterogeneity in G+C content, with the second most important trend linked with optimal growth temperature.

6.4.1 A primary trend associated with G+C content

There has been a long standing debate of the potential selective benefit of G+C ending codons at high growth temperatures (BERNARDI and BERNARDI 1986; GALTIER and LOBRY 1997; HURST and MERCHANT 2001; MUSTO *et al.* 2004; MUSTO *et al.* 2006). The wide range of optimal growth temperatures observed among Archaea provide an excellent system to investigate these possible effects, with these data including a much larger number of thermophilic species than in previous studies. From the trends in codon usage observed among Archaea, there is no evidence that natural selection has shaped the G+C content of species in a temperature-dependent manner since (i) the primary trend in GC3s is distinct and orthogonal to the secondary trend associated with growth temperature, and (ii) there is no relationship of optimal growth temperature with G+C content.

It has been noted among bacterial genomes that (for reasons which remain unknown) G+C content varies with species' oxidative tolerance; aerobes are more G+C rich than anaerobes (NAYA *et al.* 2002). Among the Archaea investigated here, there appears to be no such relationship although it remains unclear whether this reflects an issue of statistical power (with just 14 aerobes), or a genuine difference between Bacteria and Archaea. If the former is true, then the absence of any effect in these data implies the effect is small and thus relatively unimportant at explaining variation in G+C content among species.

Whilst neither optimal growth temperature nor oxygen metabolism can explain variation in GC3s, it seems that there is a simpler explanation. Consistent with observations among bacterial genomes (CHEN *et al.* 2004), (i) most heterogeneity in G+C content among the genomes of Archaea is at third codon positions, and (ii) heterogeneity in G+C content accounts for a greater proportion of variation in codon usage among genomes (67%) than within genomes (Chapter 4). These observations are consistent with the occurrence of species-specific mutational biases which can most simply explain heterogeneity in G+C content among genomes; but see Hershberg & Petrov (2010) and Hildebrand *et al*. (2010).

6.4.2 A secondary trend associated with optimal growth temperature

Previous analyses have attributed trends in codon usage bias associated with optimal growth temperature to natural selection, but in the absence of evidence (LYNN *et al.* 2002; PUIGBO *et al.* 2008). Here, optimal growth temperature was associated with a secondary trend

for which variation in the arginine AGG codon contributed most substantially. To investigate any potential selective benefit of the AGG codon among thermophilic species, the strength of selection upon arginine AGG codons was examined (S_{AGG}). Values of S_{AGG} were correlated with the secondary trend, indicating a substantial contribution by the single codon. Importantly, values of S_{AGG} were negative for the vast majority of species, indicating that the codon is selectively disfavoured and thus an unlikely candidate for thermophilic adaptation. The relationship of S_{AGG} with optimal growth temperature was relatively weak (Figure 6-6), and largely caused by a subset of species living at moderate temperatures where AGG appears to be strongly disfavoured. It is unclear why the codon should only be disfavoured at moderate temperatures. Whilst translational selection has been shown to be weaker at lower temperatures (Chapter 4), it is associated with a fourth trend, orthogonal and distinct from this temperature-associated effect. An avoidance of arginine AGG is observed among enterobacterial species and has been speculatively attributed to out-of-frame stop codon avoidance (MAYNARD SMITH and SMITH 1996), however, it is not obvious that this should be less important at high temperatures. In *E. coli*, consecutive AGG codons have been associated with a high (50%) rate of ribosomal frameshifting (SPANJAARD and VAN DUIN 1988), and this has been shown experimentally to reflect the low *in vivo* concentrations of the Arg^{UCU} cognate tRNA species (SPANJAARD *et al.* 1990). Again, there is no reason to expect this effect to be less relevant at high temperatures, and no differential patterns of tRNA usage between mesophiles and thermophiles were observed among these species of Archaea.

Alternatively, there may be no selection and the secondary trend might reflect a temperature-dependent mutational bias, and these data provide evidence that this is the more likely explanation. The correlation of optimal growth temperature with the average genome-wide excess of G over C (G-C skew) among genes suggests that optimal growth temperature impacts upon mutational processes. Furthermore, the relationship of genome-wide G-C skew with the secondary trend indicates that the trend is (in part at least) explained by G-C skew. Whilst the majority of intraspecific variation in G-C skew is typically caused by mutational biases associated with DNA replication, skews can also reflect mutational asymmetries associated with transcription (KLASSON and ANDERSSON 2006; NECSULEA and LOBRY 2007). In the situation where there are equal numbers of genes on leading and lagging strands (approximately true for most species), then the average genome-wide G-C skew value is expected to correspond to the skew due to transcription-related

biases (LOBRY and SUEOKA 2002). Transcription-associated mutational biases are expected to impact more greatly upon more highly expressed genes since they are transcribed more frequently. As expected if the secondary axis reflects transcription-associated biases, the axis leads to clearer separation of genomes in analyses of codon usage across highly expressed genes than of codon usage genome-wide (Figure 6-1). A transcription-associated mutational bias also has the potential to explain non-zero values of S_{AGG}. In this context, S_{AGG} values indicate the difference in arginine AGG codon usage between highly expressed genes and other genes. Since most S_{AGG} values are negative, this implies that transcription-associated mutations are biased away from arginine AGG towards other arginine codons on the coding strand, and are weaker at high temperatures (Figure 6-6).

Among bacterial genomes, transcription-related mutational biases have been speculatively attributed to transcription-coupled DNA repair (FRANCINO *et al.* 1996). Transcription-coupling of DNA repair is common among Bacteria and has been shown in *Escherichia coli* to result in an excess of G → C transversions on the coding strand (OLLER *et al.* 1992), and is thus expected to produce a deficit of G-ending codons such as AGG. A lack of such repair among species living at high temperatures has the potential to explain patterns of codon usage observed here among Archaea, and a notable absence of transcription-coupled repair has been observed in the thermophilic archaeon *Sulfolobus solfataricus* (ROMANO *et al.* 2007). However the mutational spectrum is expected to vary in numerous ways with growth temperature. For instance, rates of cytosine-deamination increase with temperature (LINDAHL and NYBERG 1974), and a unique DNA repair mechanism is predicted among thermophiles (MAKAROVA *et al.* 2002). So the exact molecular cause(s) of this temperature-dependent transcription-related mutational bias remain unclear.

6.4.3 Other trends

Tertiary and quaternary trends, which have not yet been identified in analyses of bacterial genomes, were identified in these analyses. The tertiary trend was associated with the average genome-wide excess of T over A among genes. It remains unclear whether this trend is unique to the archaeal domain, or if other studies have failed to identify its presence mong Bacteria.

The quaternary trend reflects variation in the strength of selected codon usage bias among species and was only present in among-species comparisons of highly expressed gene codon

usage. Given the substantial variation in the strength of selected codon usage bias among species of Archaea (Chapter 4), it might seem surprising that these analyses have revealed variability in selected codon usage bias to explain such a low proportion (4%) of variation in highly expressed codon usage among species. However these analyses only identify consistent trends among species, and are thus unable to detect much of the variation in selected codon usage bias which varies, with identity of optimal codons.

7. THE IDENTITY AND DIVERGENCE OF OPTIMAL CODONS IN ARCHAEA

7.1 Introduction

Selected codon usage bias varies in its direction among species. Differences among species in the identity of optimal codons have long been reported. For instance, early studies revealed the CUG codon is optimal for the amino acid leucine in the bacterium *Escherichia coli* (IKEMURA 1981), and yet UUG is optimal for this amino acid in the yeast *Saccharomyces cerevisiae* (BENNETZEN and HALL 1982; IKEMURA 1985). Differences in optimal codon identity emerged within the bacterial domain; the optimal codon for leucine is different again in *Bacillus subtilis* where UUA is optimal (KANAYA *et al.* 1999; SHIELDS and SHARP 1987). Although, it seems that not every amino acid is subject to optimal codon divergence; the four amino acids Phe, Ile, Tyr and Asn are each decoded by U and C-ending codons, and in each case it is the C-ending codon which is always optimal across 80 bacterial species (SHARP *et al.* 2005). Comprehensive identification of optimal codons across 160 bacterial species revealed that whilst optimal codons for U and C-ending two-fold degenerate families are generally invariant, those for other codon families can vary with mutational bias and tRNA gene content (HENRY 2007). However, it is not known whether optimal codon divergence is driven predominantly by directional mutational pressure, or following the relaxation of selected codon usage bias (SHIELDS 1990).

In Chapter 5 I examined how one aspect of selected codon usage bias, its strength, varies among Archaea. Here, I examine its direction i.e. how the identity of optimal codons varies among species with respect to base composition and tRNA gene content. Inter-family changes in optimal codon identity and tRNA gene content were investigated in their phylogenetic context, and the impact of genome-wide G+C content upon optimal codon identity and tRNA gene content was assessed. The identification of optimal codons allows for the strength of selection to be estimated for each amino acid group. In chapter three, a greater intensity of selected codon usage bias upon two-fold relative to four-fold degenerate amino acid groups was observed in a single species. Here, I extend those analyses, examining the same phenomenon in 67 species of Archaea and find it to be ubiquitous. I test the hypothesis that selection is reduced among four-fold degenerate sites due to conflicting selective effects of multiple cognate tRNA species by comparing the selection intensity for

species with either one or two forms of tRNA across two-fold degenerate sites. In direct conflict with predictions, selection is strongest among species with two forms of tRNA. Finally, by comparing sites which have and have not been subject to optimal codon divergence, I test the hypothesis that selection is reduced among four-fold degenerate sites due to the ancestral relaxation of selected codon usage bias associated with optimal codon divergence.

7.2 Methods

7.2.1 Identifying optimal codons

The same datasets of synonymous codon usage (i) genome-wide and (ii) across highly expressed genes for each of 67 species of Archaea were used as are defined in Chapters 4 and 5. Optimal codons were identified as those significantly over-represented in a dataset of highly expressed genes by comparison with the genome-wide codon usage by chi-squared tests with sequential Bonferroni correction (HENRY and SHARP 2007). Under this criterion occasionally more than one codon is optimal for a given amino acid. In these instances amino acids might by weighted disproportionately in comparative analyses. To avoid such bias, I implemented a further criterion that optimal codons must also exhibit the maximum value of the strength of selected codon usage bias (*S*) for their amino acid group. Values of *S* were computed for each synonymous codon by comparing their frequencies (i) among highly expressed genes and (ii) genome-wide (see 2.3). Resulting *S* values for the majority of synonymous codons are negative, indicating that they are underrepresented amongst highly expressed genes. Eighteen species not exhibiting selected codon usage bias (in Chapter 5) were excluded from further analyses, leaving 49 species in the final dataset.

7.2.2 Exploring variation in optimal codon identity

The numbers and anticodon identity of tRNA genes were obtained from the tRNA-scan server database (LOWE and EDDY 1997). Genome-wide G+C content across synonymously variable third codon positions (GC3s) was averaged across all protein coding genes for each of 49 archaeal species. Optimal codons and tRNA genes were considered in their phylogenetic context using the tree estimated in Chapter 5. As 42/49 species were Euryarchaeota, the other species added little information and so were excluded from

phylogenetic analyses. The vast majority of changes in optimal codon identity and tRNA gene content were species-specific and potentially subject to rapid turnover. To consider only major changes which have impacted upon the evolution of codon usage within Archaea, only large-scale inter-family changes were mapped to the phylogeny, where parsimony was used to infer ancestral states.

To explore variation in optimal codon identity with genomic G+C content, the base composition of optimal codons at third codon positions was calculated as the sum of G+C ending optimal codons minus the sum of A+T ending optimal codons assuming no base modification. The G+C content across first anticodon positions was calculated in a similar manner, as the sum of G+C first anticodon positions minus the sum of A+T first anticodon position. Phylogenetic independent contrasts were used to examine the relationship of changes in codon and anticodon G+C content with changes in genome-wide GC3s (FELSENSTEIN 1985) across 49 species of Archaea.

7.2.3 Estimating *S* for different classes of synonymous codon

Values of *S* were estimated across two (*S*2), four (*S*4) and six-fold (*S*6) degenerate amino acid groups by taking the average values of *S* across optimal codons within these degeneracy classes (Appendix G). Similarly, values of *S* were estimated separately for U and C-ending ($S2_{UC}$) and A and G-ending ($S2_{AG}$) two-fold degenerate sites by taking their average *S* values across optimal codons. There are two issues with this method. First, this method is likely to systematically upwardly bias estimates of *S* because a certain proportion of optimal codons will not be identified by chance, and these optimal codons are likely to be those which produce the lowest estimates of *S*. Nevertheless, estimates of *S* are still comparable among species since the same method has been applied to all estimates of *S*. Second, in the few cases where there are multiple optimal codons for a particular amino acid group (i.e. the second fittest codon is of high fitness), *S* may be underestimated because only one optimal codon for each amino acid group is considered. There were very few occurances of multiple optimal codons and so the impact of this effect is likely to be minimal.

To investigate the effects of multiple tRNA species upon patterns of selected codon usage bias, datasets of species with (i) one tRNA anticodon and (ii) two tRNA anticodons decoding

all A and G-ending two-fold degenerate sites were compared. There were only eight species with one tRNA anticodon decoding each of the three A and G-ending two-fold degenerate codons (Gln, Glu, Lys). To control for selected codon usage bias acting irrespective of tRNA gene content, a dataset of nine species with two tRNA anticodons decoding A and G-ending codons was selected with the criterion that species exhibited the same average *S* value (1.11) as the eight species with single tRNA genes. To investigate the impact of optimal divergence upon values of *S*, values were compared across optimal codons which had, or had not, been subject to divergence. An optimal codon divergence event involving the amino acids glycine and leucine was identified in the common ancestor of the Halobacteriales (see Results). Therefore values of *S* across optimal codons for the amino acids glycine (S_{gly}) and leucine (S_{leu}) were compared for species of the Halobacteriales with those in the sister taxa the Methanosarcinales and Methanomicrobiales.

7.3 Results

7.3.1 The identity of optimal codons in Archaea

A total of 864 putatively optimal codons were identified across 67 species of Archaea, ranging from 4 to 18 with an average of ~13 optimal codons per species (Appendix E). Since the method of identification employed here inflicts an upper limit of one putative optimal codon per amino acid, the occurrence of 18 putatively optimal codons indicates species where optimal codons were identified for all degenerate amino acid groups. Low but non-zero numbers of putatively optimal codons (average ~9) were identified for 18 species shown to lack selected codon usage bias in Chapter 5. Species lacking selected codon usage bias are not expected to possess optimal codons and so their identification for these species was unexpected. Greater numbers of optimal codons were identified across the other 49 species (average ~14), where numbers in each species were positively correlated with the strength of selected codon usage bias (*S* as in Chapter 5; $r = 0.60$, $p < 0.01$).

Among species lacking selected codon usage bias, the largest number of putative optimal codons (15) was identified for *Pyrobaculum islandicum*, a species with the third most negative *S* value ($S = -0.248$) and which exhibits the largest observed to expected standard deviation of GC3s (ratio = 3.94) among 67 species of Archaea. In this species, the distribution of GC3s values among genes appears to be bimodal; most genes including those highly expressed in *P. islandicum* exhibit GC3s values centred around 40% but a large minority (around 25%) of

genes have much higher values centred around 70%. Thus, the large minority of G+C rich genes impacts upon genome-wide codon usage, and so contrasting genome-wide and highly expressed gene codon usage reveals an excess of A and U-ending codons across highly expressed genes, which are therefore identified as putatively optimal. While it is not clear what has caused the G+C rich codon usage in a minority of genes in *P. islandicum*, it cannot be translational selection, since selection impacts upon the codon usage of highly expressed genes. So in this case putatively identified optimal codons are artifacts. This example illustrates problems that can arise from automated methods of optimal codon detection. To avoid erroneous optimal codons impacting upon these analyses, the 18 species which did not exhibit selected codon usage bias for four amino acids (*S*; Chapter 5) were excluded from subsequent analyses.

7.3.2 Variation in optimal codons among Archaea

The identity of optimal codons varies for some amino acid groups but not others. For U and C-ending two-fold degenerate amino acid groups, the optimal codon appears to be invariantly C-ending (Table 7-1). One exception is the amino acid cystiene, for which optimal codons are uncommon and generally U-ending. The only three-fold degenerate amino acid, isoleucine, behaves similarly to the U and C-ending two-fold degenerate amino acids, with the C-ending codon optimal in the vast majority of species. The situation is more complex for the A and G-ending two-fold degenerate amino acid groups (Gln, Glu, Lys). For the majority of species the G-ending codon is optimal, but in a few instances, largely restricted to the Methanobacteriales and Methanococcales, it is the A-ending codon which is optimal. Four and six-fold degenerate amino acids are highly variable in their optimal codon identity across all archaeal species.

N	Number of N-ending optimal codons among Archaea																	
	Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
C	46	42	1	36	46	42	35	--	--	--	7	9	7	24	25	3	24	14
T	0	0	10	0	0	0	1	--	--	--	10	37	6	0	5	18	3	2
A	--	--	--	--	--	--	4	4	8	4	22	2	11	8	6	15	7	8
G	--	--	--	--	--	--	--	39	17	41	0	0	12	2	1	7	11	1

Table 7-1 Number of N-ending optimal codons among Archaea

Summed optimal codons for each amino acid across 49 species of Archaea.

Many changes in optimal codon identity appear to be sporadic, occurring within families, and may be transient in nature. To investigate large-scale changes in optimal codon identity which have been influential in the evolution of codon usage bias, changes in optimal codon identity between archaeal families were considered in their phylogenetic context. From this, three major changes in optimal codon identity among families are apparent (Table 7-2). First, there appears to have been gain of selected codon usage bias in the ancestor of the Methanosarcinales, Methanomicrobiales, and Halobacteriales across several amino acid groups: Glu, Lys, Ala, Gly, Pro, Leu and Ser. Second, there are changes in optimal codon identity in the common ancestor of the Halobacteriales for the amino acids glycine and leucine. Whilst the change for leucine does not impact upon the G+C content of the optimal codon, the change for glycine (GGU → GGC) is towards a more G+C rich optimal codon, and it is noteworthy that the Halobacteriales are the most G+C rich of all archaeal clades, with an average genome-wide GC3s of 0.77 across species. Finally, changes in optimal codon identity from G to A-ending optimal codons across two-fold degenerate families appear to have occurred in the common ancestor of the Methanobacteriales. For the amino acid lysine, the change may have occurred less recently, in the shared ancestor of the Methanococcales and Methanobacteriales. Both families are the most A+T rich of Archaea, with a combined average genome-wide GC3s value of 0.25 across species. Each of the changes in optimal codon identity has led to more A+T rich optimal codons in these species. Thus there appears to be a potential link between changes in optimal codon identity and changes in genome-wide G+C content.

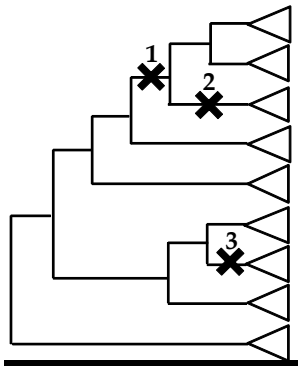
Cladogram	Family	N	GC3s	Ancestral optimal codon																	
				Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
	Methanosarcinales	6	0.49	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU		ACC	GUC	CGU	CUC	UCC
	Methanomicrobiales	5	0.64	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCG	ACC	GUC	CGU	CUC	UCC
	Halobacteriales	8	0.77	AAC	GAC	--	--	UUC	UAC	AUC	CAG	--	AAG	GCA	GGC		ACC	GUC	CGU	CUG	UCC
	Archaeoglobales	3	0.49	AAC	GAC	--	--	UUC	UAC	AUC	CAG	GAG	AAG	--	GGU	CCG	--	--	AGA	CUC	
	Thermoplasmatales	3	0.40	AAC	--	--	CAC	UUC	UAC	--	CAG		AAG	--	GGU	--	--	--	CGU		--
	Methanococcales	6	0.22	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--		GCU	GGU	CCU	ACA	GUU	AGA		UCA
	Methanobacteriales	4	0.28	AAC	GAC	--	CAC	UUC	UAC	AUC	CAA	GAA	AAA		GGU	CCU		GUA		CUC	UCU
	Methanopyrales	1	0.74	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCC	GGU	CCG	ACC	GUG		CUG	AGC
	Thermococcales	6	0.51	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCU	GGU	CCA	ACC	--	AGA	CUC	AGC

Table 7-2 Changes in optimal codon identity among families of the Euryarchaeota

Ancestral optimal codon states were inferred for 9 euryarchaeote families with N species by parsimony. GC3s indicates genome-wide averages across families. The instances where no optimal codon is ancestral are indicated by - - . Spaces indicate where there was insufficient information to infer states and colours indicate the base at the third optimal codon position. The phylogenetic tree topology is taken from Figure 5-1 in Chapter 5. Three events in optimal codon divergence are indicated on the cladogram: (1) Gain of optimal codons for Ala, Thr, Val and Ser; (2) Change in Gly from GGU to GGC and Leu from CUC to CUG; (3) Changes in Gln to CAA, and Glu to GAA in the Methanobacteriales. It is unclear if the change to Lys AAA occurred on the branch to the Methanobacteriales or in their shared common ancestor with the Methanococcales.

To investigate the relationship of genome-wide G+C content and optimal codon identity, the G+C content across third optimal codon position was considered across four categories of synonymous codons: (i) two-fold degenerate U and C-ending, (ii) two-fold degenerate A and G-ending, (iii) four-fold degenerate, and (iv) six-fold degenerate groups. As expected if U and C-ending two-fold degenerate sites always contain C-ending optimal codons, the number of C minus U-ending optimal codons in each archaeal species is always positive, and is not related to genome-wide GC3s among species (Figure 7-1A). The situation is different for G and A-ending two-fold degenerate amino acid groups, where the number of G minus A-ending optimal codons can be positive or negative, indicating that either can be optimal (Figure 7-1B). There are a much larger number of positive than negative values, indicating that G-ending codons are predominantly optimal across these amino acids for most species. Few species possess the intermediate value of zero, and the few species with negative values where A-ending optimal codons are prevalent are restricted to the A+T rich families, the Methanobacteriales and Methanococcales. Across both four and six-fold degenerate sites, there are positive correlations with the numbers of (G+C) – (A+T) ending optimal codons and average genome-wide GC3s among genes ($r = 0.79$, $p < 0.001$, Figure 7-1C and $r = 0.63$, $p < 0.001$, Figure 7-1D respectively), which are robust to correction for phylogenetic non-independence ($r = 0.69$, $p < 0.001$; $r = 0.47$, $p < 0.001$ respectively).

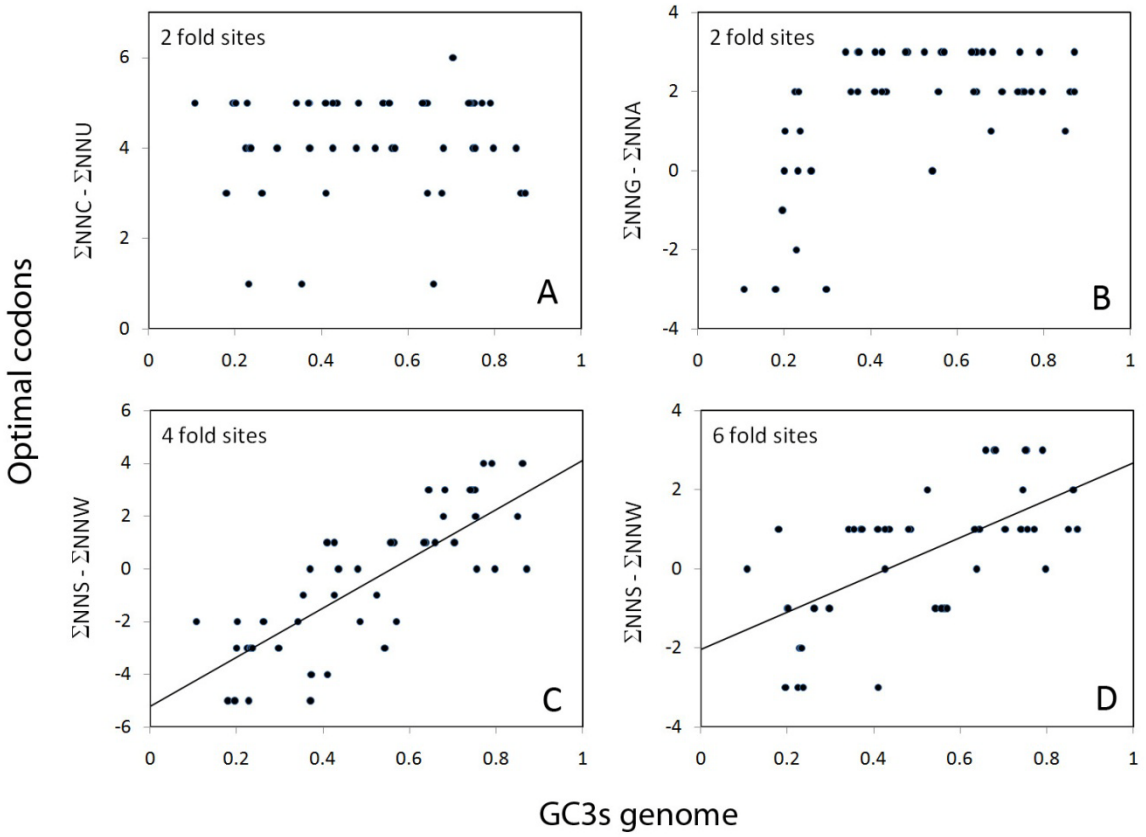


Figure 7-1 Variation in G+C content at third optimal codon position with genome-wide GC3s among Archaea

Third optimal codon position G+C content is scored as the sum of G+C (S) -ending minus the sum of A+U (W) -ending optimal codons and was assessed for: (A) C and U-ending two-fold degenerate, (B) A and G-ending two-fold degenerate, (C) four-fold degenerate, and (d) six-fold degenerate amino acid groups in 49 species of Archaea.

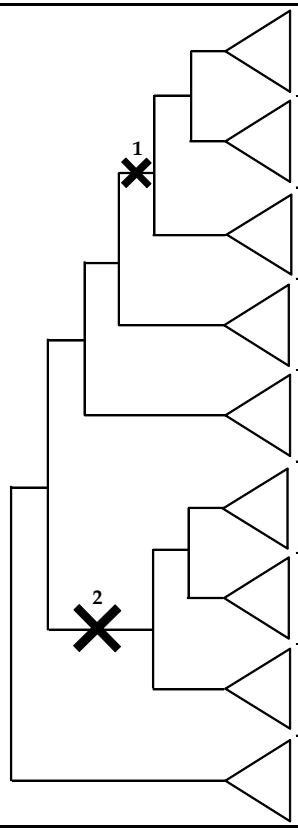
7.3.3 Variation in tRNA gene content among Archaea

The identity of optimal codons is expected to coevolve with tRNA abundance (BULMER 1991a) and abundances appear to be largely determined by tRNA gene copy number (KANAYA *et al.* 1999). Therefore distributions of tRNA gene contents were explored for Archaea. Among the 49 species of Archaea investigated here, numbers of tRNA genes vary across a restricted range (31-63) by comparison with numbers across 80 bacterial genomes (28-126 (SHARP *et al.* 2005)). The lowest numbers of tRNA genes are observed across the sister taxa, the Methanococcales and Methanobacteriales, which range from 34 to 50 with a mean of 38.4. These species also contain minimal numbers of different anticodons, ranging from 31to 35. Most Archaea (29/45), by contrast possess at least 43 different anticodons and few redundant tRNA genes, with nearly 30% of species (14/45) containing either 46 or 47 tRNA genes (Appendix F).

To determine if changes in tRNA gene content coincide with changes in optimal codon usage, inter-family changes were mapped to the archaeal phylogeny (Table 7-3). From this, a duplication of the alanine TGC tRNA is apparent in the common ancestor of the Methanomicrobiales, Methanosarcinales and Halobacteriales. This event coincides with the gain of selected codon usage bias for the complementary alanine GCA codon (Table 7-2). However gain of selected codon bias was also apparent for the amino acids threonine, valine and serine, and yet these events do not appear to have been accompanied by changes in tRNA gene content. Similarly, there are no apparent changes in tRNA gene content that can be associated with optimal codon divergence observed for leucine or glycine in the Halobacteriales, meaning that in total there are more independent than concerted changes in optimal codon identity and tRNA genes that have been identified.

Strikingly, large-scale deletion of many tRNA genes with C at their first anticodon position appears to have occurred for the amino acids: Glu, Lys, Ala, Gly, Pro, Leu and Ser in the common ancestor of the Methanococcales, Methanobacteriales and Methanopyrales. Futher deletions for C-starting tRNAs might also have occurred in this lineage for the amino acids Gln, Thr, Val and Arg since they are present in only one of the Methanococcales and Methanobacteriales sister taxa. In any case there have been tRNA deletions across all amino acids for which optimal codon identity is subject to change at some point in these families. Deletion events in the common ancestor of Methanococcales, Methanobacteriales and

Methanopyrales do not directly coincide with mapped changes in optimal codon identity or mutational bias (Table 7-2). The simplest way for many tRNA genes to be deleted simultaneously would be via a single event involving the loss of a single tRNA gene operon. Examination of the genomic context of tRNA genes in other species with C-starting anticodons reveals that these genes are distributed broadly across the chromosome. If these contexts are indicative of the ancestral state, it suggests that the tRNA genes were deleted on multiple occasions rather than as a single event in an operon deletion.



Cladogram	Family	Sp[a]	N[b]	Ancestral number of N-starting anticodons																
				Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
	Methanosarcinales	6	C	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	2	2	1
			T	--	--	--	--	--	--	0	1	1	1	2	1	1	1	1	2	2	1
			G	1	1		1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Methanomicrobiales	5	C	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	2	2	1
			T	--	--	--	--	--	--	0	1	1	1	2	1	1	1	1	2	2	1
			G	1	1		1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Halobacteriales	8	C	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	2	2	1
			T	--	--	--	--	--	--	0	1	1	1	2	1	1	1	1	2	2	1
			G	1	2	1	1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Archaeoglobales	3	C	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	1	2	1
			T	--	--	--	--	--	--	0	1	1	1	1	1	1	1	1	1	2	1
			G	1	1	1	1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Thermoplasmatales	3	C	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	2	2	1
			T	--	--	--	--	--	--	0	1	1	1	1	1	1	1	1	2	2	1
			G	1	1	1	1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Methanococcales	6	C	--	--	--	--	--	--	--	0	0	0	0	0	0	0	1	0	0	0
			T	--	--	--	--	--	--	0	1	2		1	1	1	1	1	2	2	1
			G	1		1	1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Methanobacteriales	4	C	--	--	--	--	--	--	--	1	0	0	0	0	0	1	0	1	0	0
			T	--	--	--	--	--	--	0	1	1	1		1	1	1	1	2	2	1
			G	1	1	1	1	1	1	1	--	--	--	1	1	0	1	1	1	1	2
	Methanopyrales	1	C	--	--	--	--	--	--	--	0	0	0	0	0	0	0	0	0	0	0
			T	--	--	--	--	--	--	0	1	1	1	1	1	1	1	1	2	2	1
			G	1	1	1	1	1	1	1	--	--	--	1	1	1	1	1	1	1	2
	Thermococcales	6	C	--	--	--	--	--	--	--	1	1	1	1	1	1	1	1	2	2	1
			T	--	--	--	--	--	--	0	1	1	1	1	1	1	1	1	2	2	1
			G	1	1	1	1	1	1	1	--	--	--	1	1	1	1	1	1	1	2

Table 7-3 Distribution of anticodons among the Euryarchaeota

[a]The number of species in each family [b]The base at the first anticodon position

Ancestral genome-wide numbers of tRNA gene anticodons inferred for 9 euryarchaeote families across 46 species by parsimony. Spaces indicate where it was not possible to identify the ancestral state. Changes in tRNA content are indicated with crosses: (1) Gain of Ala TCG anticodon. (2) Large scale loss of many tRNA genes with C at the first anticodon position.

In the previous section I showed that genome-wide G+C content is linked with optimal codon identity. Here, the relationship of genome-wide G+C content with G+C content across first anticodon positions was investigated across four amino acid classes (as in Figure 7-1). Across tRNAs decoding the U and C-ending two-fold degenerate codons, there is only one (G-starting) tRNA anticodon, and its copy numbers are unrelated to GC3s (Figure 7-2A). For most (27) archaeal species, the number of C minus U-starting anticodons (which decode the A and G-ending amino acids) is zero, with 26 species of Archaea containing single copies for each possible C or U-starting tRNA anticodon (Figure 7-2B). However in a minority of A+T rich species (the Methanococcales and Methanobacteriales), C-starting tRNAs are absent, giving rise to a trend of genomic G+C content with anticodon G+C content. A clear outlier in this trend is the methanopyrale *Methanopyrus kandleri*, which (as noted) lacks the C-starting anticodons and yet exhibits a G+C rich genome (GC3s = 0.74). Across four-fold and six-fold degenerate amino acids, the independent contrats numbers of (G+C) – (A+U) starting anticodons is positively correlated with genome-wide GC3s (Figure 7-2C, $r = 0.45$, $p < 0.01$; Figure 7-2D, $r = 0.54$, $p < 0.01$ respectively)., which remain robust to correction for phylogenetic non-independence ($r = 0.25$, $p = 0.04$; $r = 0.42$, $p < 0.01$ respectively).

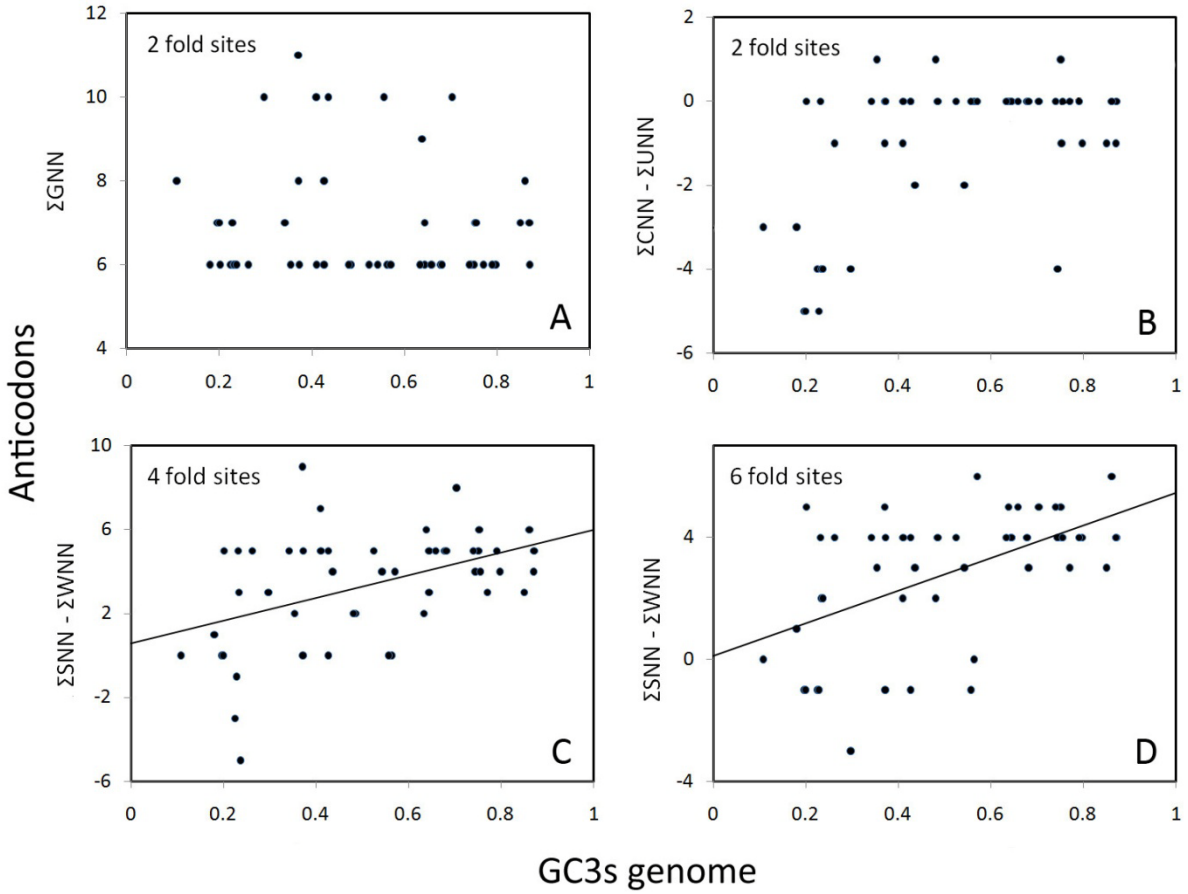


Figure 7-2 Variation in G+C content at first anticodon position with genome-wide GC3s among Archaea

First anticodon position G+C content is scored as the sum of G+C-ending minus the sum of A+U-ending tRNA anticodons across four amino acid classes (see Figure 7-1) in 49 species of Archaea.

7.3.4 The strength of selected codon usage bias (*S*) across two and four-fold degenerate amino acid groups in Archaea

The identification of optimal codons allows for the estimation of the strength of selection (*S*) acting upon them (Appendix G). In Chapter 3, selected codon usage bias was found to be more intense across two than across four-fold degenerate amino acid groups. Here, *S* was estimated across two (*S*2), four (*S*4), and six-fold (*S*6) degenerate amino acid groups (Figure 7-3). Among all 49 species of Archaea in this dataset, the strength of selected codon usage bias across two-fold degenerate amino acids was greater than across four-fold degenerate amino acids, and the distributions of *S*2 and *S*4 were found to differ significantly (paired t-test $p < 0.001$). In most (35) species, *S*6 was greater than *S*4 but less than *S*2, and the distributions of *S*6 and *S*4, and *S*6 and *S*2 also differ from one another (paired t-tests both $p < 0.001$).

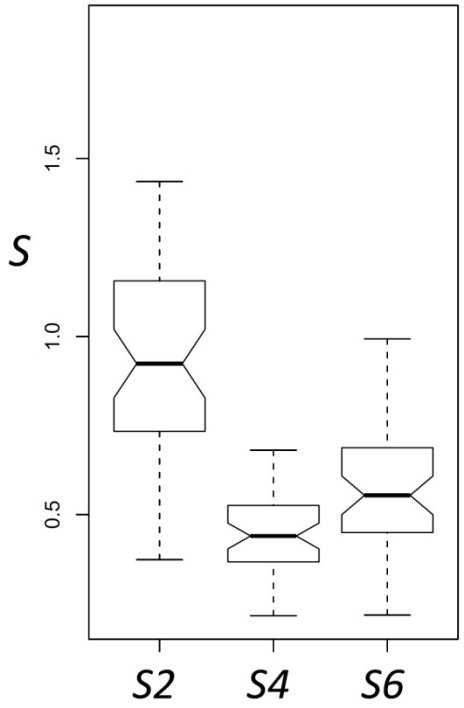


Figure 7-3 The distribution of *S* values across two, four and six-fold degenerate amino acid groups among 49 species of Archaea

Dark horizontal lines indicate median values; boxed areas indicate the 50% range of values; error bars indicate the 95% range of values.

It is possible that the distributions differ due to a systematic bias in the method since the *S* values were obtained using a two-state model (BULMER 1991b). This means that in the situation where there are more than two synonymous codons (i.e. across four and six-fold degenerate sites), codons were binned into two allelic classes of optimal and non-optimal codons. Thus any variation in the fitness of non-optimal codons may lead to inaccuracies in the estimation of *S*. However, any systematic error seems incapable of explaining the values obtained here. If the observation that $S2 > S4$ reflects a systematic underestimation of *S*, then the problem should be worse among six-fold degenerate sites and yet $S6 > S4$.

Values of *S* may also be estimated for U+C ($S2_{UC}$) and A+G-ending ($S2_{AG}$) two-fold degenerate amino acid groups separately. To assess any impact of the number of different forms of tRNA upon the strength of selected codon usage bias, values of $S2_{AG}$ were compared for species with either one or two different tRNA anticodons decoding A and G-ending two-fold degenerate sites (Table 7-3).

Species	tRNA anticodons						$S2_{UC}$	$S2_{AG}$
	Gln		Glu		Lys			
	TTG	CTG	TTC	CTC	TTT	CTT		
Methanobrevibacter smithii	1		1		1		0.92	0.87
Methanocaldococcus jannaschii	1		2		1		1.27	0.72
Methanococcus aeolicus	1		2		2		1.02	0.75
Methanococcus maripaludis S2	1		2		2		1.81	0.73
Methanococcus vannielii	1		2		2		1.44	0.67
Methanopyrus kandleri	1		2		1		1.19	1.08
Methanocaldococcus fervens	1		2		1		1.26	0.46
Methanocaldococcus vulcanius	1		2		1		0.75	0.76
Methanococcoides burtonii	1	1	1	1	1	1	0.97	1.39
Halobacterium salinarum	1	1	1	1	1	1	1.57	1.11
Thermococcus onnurineus	1	1	1	1	1	1	1.23	1.58
Haloarcula marismortui	1	1	1	1	2	1	1.52	1.42
Ignicoccus hospitalis	1	1	1	1	1	2	1.07	1.31
Rice Cluster I MRE50	1	1	1	1	1	1	1.38	1.68
Methanocella paludicola	1	1	1	1	1	1	1.32	1.44
Natronomonas pharaonis	1	1	1	1	1	1	1.40	1.33
Halorhabdus utahensis	1	1	1	1	1	1	1.30	1.06

Table 7-3 The strength of selected codon usage bias in species of Archaea with alternative patterns of tRNA anticodons

The strength of selected codon usage bias (S) is given for U and C-ending ($S2_{UC}$), and A and G-ending ($S2_{AG}$) two-fold degenerate sites. Anticodons indicated decode the A and G-ending codons.

While the strength of selected codon usage bias is no different across U and C-ending two-fold degenerate codons ($S2_{UC}$) for species containing either one (n = 8) or two (n = 9) forms of tRNA anticodon (t-test, p = 0.48), it is reduced across A and G-ending two-fold degenerate sites ($S2_{AG}$) for species with single tRNA anticodons ($p < 0.001$). Values of $S2_{AG}$ are significantly lower than values of $S2_{UC}$ for species with single tRNA forms (paired t-test, p = 0.02), but not for species with two tRNA forms (p = 0.54).

To compare the strength of selected codon usage bias across optimal codons which have been subject to divergence with those which have not, the strength of selected codon usage bias (S) was estimated for the amino acids glycine (S_{gly}) and leucine (S_{leu}), which have been subject to optimal codon identity divergence in the Halobacteriales. These values of S across the Halobacteriales were compared with values across their closest relatives, the Methanosarcinales and Methanomicrobiales, where optimal codon identity has not diverged for these amino acids (Table 7-4).

Species	S	S_{gly}	S_{leu}
Halobacteriales			
Haloarcula marismortui	1.032	0.47	0.21
Halobacterium salinarum	0.959	0.30	0.11
Halorubrum lacusprofundi	0.536	0.44	0.29
Halomicrobium mukohataei	1.168	0.27	0.27
Halorhabdus utahensis	1.430	0.35	0.38
Haloterrigena turkmenica	0.925	0.61	0.21
Natrialba magadii	1.081	0.38	0.42
Natronomonas pharaonis	1.254	0.29	0.32
Methanosarcinales			
Methanococcoides burtonii	0.910	0.70	0.65
Methanosaeta thermophila PT	0.665	0.19	0.09
Methanosarcina acetivorans	0.852	0.61	0.47
Methanosarcina barkeri fusaro	0.671	0.38	0.41
Methanosarcina mazei	0.892	0.73	0.46
Rice Cluster I MRE50	1.058	0.72	0.69
Methanomicrobiales			
Methanocorpusculum labreanum	0.884	0.99	0.39
Methanoculleus marisnigri	0.470	0.83	0.42
Methanoregula boonei	0.690	0.58	0.45
Methanospirillum hungatei	0.628	0.69	0.31
Methanocella paludicola	1.241	0.93	0.35

Table 7-4 The strength of selected codon usage bias across optimal codons which have and have not been subject to divergence

The strength of selected codon usage bias (S) is given for amino acids glycine (S_{gly}), and leucine (S_{leu}) two-fold degenerate sites. The amino acids glycine and leucine were found to be subject to optimal codon divergence in the common ancestor of the Halobacteriales (Table 7-2).

Values of S across the two-fold degenerate amino acids (Phe, Ile, Tyr and Asn - obtained in Chapter 5) are on average greater for species of Halobacteriales (average S = 1.00) than those of its sister taxa, the Methanosarcinales and Methanomicrobiales (average S = 0.81). Yet values of S_{gly} and S_{leu} are on average lower for species of Halobacteriales (average S_{gly} = 0.39, S_{leu} = 0.28) than for species of Methanosarcinales and Methanomicrobiales (average S_{gly} = 0.67, S_{leu} = 0.42), and the distributions of S_{gly} and S_{leu} for the Halobacteriales differ significantly from these sister taxa (t-tests, p = 0.003 and p = 0.026 respectively).

7.4 Discussion

7.4.1 Optimal codon identity by comparison with other studies

Optimal codons for Archaea were identified in these analyses, and by comparison with analyses of bacterial genomes, reveal some consistent trends. In both domains, optimal codon identity is invariant for the U and C-ending two fold degenerate groups (HENRY 2007; SHARP *et al.* 2005). Yet for the other amino acid groups, optimal codon identity varies among species. A study of 160 bacterial species mapped large-scale changes in optimal codon identity to the bacterial phylogeny, and 16 instances were identified where optimal codon divergence coincides with changes in tRNA gene content (HENRY 2007). Here among Archaea, coordinated divergence of optimal codons and tRNA genes were less commonly detected. In one case, the gain of selected codon usage bias for the alanine GCA codon coincided with the gain of an additional complementary UGC anticodon in the common ancestor of the Methanomicrobiales, Methanosarcinales and Halobacteriales. Failure to detect coordinated events among Archaea most likely reflects the low numbers and diversity of archaeal tRNA genes, meaning that few changes in tRNA gene content are apparent. The availability of fewer archaeal genomes sequences (49) than Bacteria (160) also means that less information was available to detect divergence events spanning a similar evolutionary timescale.

Optimal codon divergence in the bacterial domain was found to coincide with directional changes in mutational bias (HENRY 2007). Consistent with these patterns, divergence in third optimal codon position G+C content was correlated with divergence of genome-wide GC3s across four and six-fold degenerate sites in Archaea. Studies of bacterial genomes have not yet considered the impact of mutational biases upon tRNA genes. Here among Archaea, changes in first anticodon position G+C content were correlated with changes in genome-

wide GC3s. This result is consistent with a role for directional mutational pressure in shaping tRNA gene content via its impact upon genome-wide codon usage, and the subsequent selective pressure this is expected to exert on tRNA gene copies.

Across the bacterial domain, instances of optimal codon divergence are generally coordinated, with the divergence of optimal codons from as many as five amino acid groups mapped to the same internal branch of the bacterial phylogeny (HENRY 2007). These divergence events are associated with changes in the numbers of tRNA genes with C at the first anticodon position. Here in Archaea, there is a striking example of coordinated deletions of tRNA genes with C at the first anticodon position for at least seven amino acids (Table 7-3), and collectively these observations point to a widespread role for the divergence of C-starting tRNAs in shaping optimal codon identities. Whilst phylogenetic mapping of these deletions was not found to directly coincide with optimal codon divergence when considering changes in optimal codon identity and tRNA gene content independently (Table 7-5A), it seems highly unlikely that seven tRNA genes were simultaneously deleted by chance. Correlations of G+C content across genome-wide synonymously variable third codon positions, G+C contents across first anticodon and third optimal codon positions suggest that divergence of all three factors are not independent of one another. In this case, patterns are most simply explained if the tRNA deletions were driven by changes in mutational bias towards A+T richness, which impact upon genome-wide codon usage, thus exerting selective pressue for changes in tRNA gene content and optimal codon identity (Table 7-5B). However, it remains unclear why natural selection has not favoured the gain of copies of C-starting tRNA anticodons in the methanopyrale *M. kandleri*.

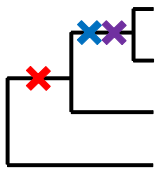
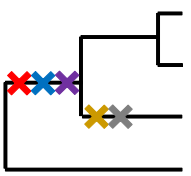
Family	tRNA genes	GC3s	Optimal codons	A	B
Methanococcales	Absent	Low	Many A+U ending		
Methanobacteriales	Absent	Low	Many A+U ending		
Methanopyrales	Absent	High	Most G+C ending		
Other Euryarchaeota	Present	Moderate	Most G+C ending		

Table 7-5 Divergence of tRNA genes, mutational bias and optimal codon identity across families of the Euryarchaota

The presence of tRNA genes for seven amino acids with C at the first anticodon position is indicated. Crosses denote divergence of traits indicated in the table with the same colour. Scenario (A) is most parsimonious where divergence of tRNA genes, average genome-wide GC3s and optimal codon identity are considered as independent, requiring a total of three events. Scenario (B) is most parsimonious where divergence of tRNA genes, GC3s and optimal codons are considered to be linked, and requires a total of two events.

A more recent study attempted to identify optimal codons in a wide range of species, including Archaea (HERSHBERG and PETROV 2009), however the method of optimal codon identification that was implemented is expected to produce erroneous results. Briefly, the method considers a codon optimal if its frequency is best correlated with a general measure of codon bias (N_c, see 2.1.5) among genes. Yet trends other than selected codon usage bias can impact upon N_c. Multivariate analyses of codon usage within archaeal genomes revealed that general patterns in codon bias are typically dominated by variation in GC3s or G-C skew, with only a few species (5) exhibiting major trends in selected codon usage among genes (Chapter 4). Since it has been demonstrated both theoretically (EHRENBERG and KURLAND 1984; WILKE and DRUMMOND 2006) and experimentally (IKEMURA 1981; IKEMURA 1985) that the highly expressed genes are the primary target of translational selection, then optimal codons may be simply identified as those overrepresented among highly expressed genes (see Methods). By comparison with this study, the method of Hershberg and Petrov appears to have incorrectly identified 159 optimal codons across 34 archaeal species exhibiting significant selected codon usage bias. Errors in optimal codon detection were greatest across four and six-fold degenerate codons (those which are subject to most optimal codon divergence), with incorrect identification in 50% or more species for the amino acids: Ala, Arg, Gly, Leu and Ser.

7.4.2 Variation in the strength of selected codon usage bias among amino acid groups in Archaea

In Chapter 3, the strength of selected codon usage bias was found to be greater across two-fold (*S*2) than four-fold (*S*4) degenerate amino acid groups for the archaeon *M. maripaludis* but not for the bacterium *Escherichia coli*. Here these analyses were extended to all Archaea exhibiting significant selected codon usage bias, and in every case *S*2 was greater than *S*4. Furthermore a subtle but significant effect was detected whereby the strength of selected codon usage bias across six-fold degenerate amino acids (*S*6) is greater than *S*4 (Figure 7-3).

While two-fold degenerate U and C-ending codons are translated by single tRNA species, two-fold degenerate A and G-ending codons can be translated by one or two tRNAs with different anticodons. Similarly, four-fold degenerate amino acids are translated by as many as three different tRNA species. It might be that alternative cognate tRNA anticodons best translate different synonymous codons, and so the presence of multiple cognate tRNA species might reduce the selective benefit of any particular codon. In Chapter 3, I speculated that the conflicting selective effects of multiple cognate tRNA species might explain the reduced magnitude of *S*4 relative to *S*2. Here, this hypothesis was tested by comparing the selection intensity across two-fold degenerate A and G-ending (S_{AG}) codons in species where these sites are decoded by either one or two tRNA species (Table 7-3). If selection is reduced due to the conflicting selective effects of multiple tRNA species then the expectation is that species with single tRNA species decoding A and G-ending sites exhibit the greatest values of S_{AG}. However, in direct conflict with this prediction, the opposite was found to be true.

Patterns of tRNA gene divergence revealed that species with single tRNA anticodons decoding two-fold degenerate A and G-ending codons are the product of a coordinated large-scale tRNA deletion event, which most likely occurred in the common ancestor of the Methanopyrales, Methanobacteriales and Methanococcales, and may have coincided with optimal codon divergence. Thus, there may be a bias in the above dataset; optimal codons may have been subject to more recent divergence in species with single tRNA anticodons than in species with pairs of tRNA anticodons. Optimal codons may diverge either, (i) driven by mutational pressure or, (ii) following the relaxation of selection (SHIELDS 1990). If optimal codon divergence is preceded by a relaxation of selected codon bias, this could explain why estimates of the long-term strength of selected codon usage bias (S_{AG}) are

weaker in species with single tRNA anticodons, which have been subject to more recent optimal codon divergence. Under this model of optimal codon divergence, selection is relaxed for certain amino acids but not others, and so is not expected to be driven by population bottlenecks which would impact upon all amino acid groups.

This hypothesis has the potential to explain why values of *S*4 and *S*6 are lower than *S*2. Optimal codon divergence predominantly impacts upon four and six-fold degenerate sites and so widespread divergence via relaxation of selected codon usage bias could mean that long-term estimates of selection generally appear weaker across these sites than across two-fold degenerate U and C-ending invariant optimal codons. To test this hypothesis, the strength of selected codon usage bias was compared for optimal codons (for the amino acids glycine and leucine) which diverged in the common ancestor of the Halobacteriales with the same codons in the sister taxa the Methanosarcinales and Methanomicrobiales (Table 7-2). Whilst the strength of selected codon usage bias (*S*) was higher across two-fold degenerate invariant optimal codons for the Halobacteriales compared with the Methanosarcinales and Methanomicrobiales, values were lower in the Halobacteriales across optimal codons for the amino acids glycine and leucine, which have been subject to divergence. Interestingly, in the recent ancestry of the bacterium *Escherichia coli*, a species where values of *S*2 were not significantly different to *S*4, only two instances of optimal divergence have been reported; one for the two-fold degenerate amino acid glutamine, and the other for the four-fold degenerate amino acid proline (HENRY 2007). In each case values of *S* for these amino acids are much lower than values of *S* across the two-fold degenerate amino acids: Phe, Ile, Tyr and Asn ($S = 1.49$, $S_{gln} = 0.84$, $S_{pro} = 0.78$). Thus a consistent pattern seems to have emerged; selection is weaker across optimal codons which have been subject to divergence, consistent with a widespread role for the relaxation of selected codon usage bias in optimal codon divergence. It will be interesting to extend these analyses to the bacterial domain where larger numbers of optimal codon divergence events have been documented.

8. EXPLORING THE STRENGTH OF SELECTED CODON USAGE BIAS FOR ACCURATE TRANSLATION IN ARCHAEA

8.1 Introduction

There has been much debate as to the nature of selected codon usage bias. Is selection for the efficiency or accuracy of translation, and if both, which is most important? Whilst selection for efficient translation is expected to occur at a similar intensity across all synonymous sites of highly expressed genes, selection for accurate translation is expected to be most intense across functionally and/or structurally important residues. Therefore optimal codon frequencies across evolutionarily conserved sites have been explored as a means of detecting the presence of selection for accurate translation in a variety of organisms, and in most cases significant associations of optimal codons with conserved sites are reported (AKASHI 1994; DRUMMOND and WILKE 2008; HARTL *et al.* 1994; STOLETZKI and EYRE-WALKER 2007).

Whilst previous analyses have tested for the presence of selected translational accuracy, none have estimated its strength. Comparative analyses of the strength of translational accuracy-selected codon usage bias might allow us to answer some interesting evolutionary questions. For instance, if the selective benefit of translational accuracy was an efficiency saving, then we might expect accuracy-selection to be most relevant to species with rapid growth rates and thus positively correlate with the strength of efficiency-selected codon usage bias. Alternatively, the opposite might be true if a trade off exists whereby the most efficient codons are not the most accurate (DIX and THOMPSON 1989; PARKER and PRECUP 1986).

A variety of observations have indicated that obtaining a precise amino acid sequence is likely to be more critical at high temperatures (DRAKE 2009; FRIEDMAN *et al.* 2004), leading to the prediction that the strength of accuracy-selected codon usage bias is greatest among thermophilic species. The estimation of the strength of accuracy-selected codon usage bias in Archaea, with wide ranging growth temperatures, may provide an ideal dataset to test this hypothesis. Here, I therefore examine codon usage bias across conserved and non-conserved amino acid residues in order to explore the estimation of the strength of accuracy-selected codon usage bias in Archaea.

8.2 Methods

8.2.1 Estimating the strength of accuracy-selected codon usage bias (S_{acc})

A classical population genetics model of codon usage (BULMER 1991b) considers the situation where an amino acid is encoded by two synonyms, with a selective difference of s, and where mutation rates from one to the other can be represented by u and v. Then, under the combined forces of selection, mutation and random genetic drift, and with an effective population size of N_e, the equilibrium frequency of the optimal codon is expected to be:

$$P = \frac{e^S V}{(e^S V + U)} \qquad (1)$$

Where $S = 2\ N_e\ s$, $U = 2\ N_e\ u$, and $V = 2\ N_e\ v$.

To disentangle the effects of efficiency- and accuracy-selected codon usage bias, I first assume that accuracy-selection applies only to all conserved amino acid residues. I extend the model, partitioning the selection coefficient, s, into accuracy (s_{acc}) and efficiency (s_{eff}) components. Since the selective benefits conferred to optimal codon usage in accuracy and efficiency are expected to act independently of one another, I assume their effects to be additive:

$$s = s_{acc} + s_{eff} \qquad (2)$$

Then, the equilibrium frequency of optimal codons across conserved sites in highly expressed genes (P_{HC}), assumed to be functionally and/or structurally important and thus under the influence of both accuracy- and efficiency-selected codon usage bias is then:

$$P_{HC} = \frac{e^{(s_{acc} + s_{eff})} V}{(e^{(s_{acc} + s_{eff})} V + U)} \qquad (3)$$

Among the non-conserved sites of highly expressed genes, where obtaining the correct amino acid sequence is less critical, accuracy-selected codon usage bias is assumed to be

weak or absent. The equilibrium frequency of optimal codons across these sites (P_{HN}) is expected to be:

$$P_{HN} = \frac{e^{(s_{eff})}V}{(e^{(s_{eff})}V+U)} \quad (4)$$

Where all forms of selected codon usage bias are absent (or so weak as to be ineffective), optimal codon frequencies (P_L) are determined simply by mutational patterns:

$$P_L = \frac{V}{(V+U)} \quad (5)$$

The strength of selection (S) corresponds to twice the product of the long term effective population size and the selection coefficient. Rearranging (3) and substituting in (5) allows for the estimation of the combined strength of accuracy- ($S_{acc} = 2N_e s_{acc}$) and efficiency- ($S_{eff} = 2N_e s_{eff}$) selected codon usage bias:

$$S = S_{acc} + S_{eff} = ln\frac{P_{HC}(1-P_L)}{(P_L(1-P_{HC}))} \quad (6)$$

Similarly, rearranging (4) and substituting in (5) allows for the estimation of the strength of efficiency-selected codon usage bias:

$$S_{eff} = ln\frac{P_{HN}(1-P_L)}{(P_L(1-P_{HN}))} \quad (7)$$

The strength of accuracy selected codon usage bias (S_{acc}) is then the difference between (6) and (7):

$$S_{acc} = ln\frac{P_{HC}(1-P_L)}{(P_L(1-P_{HC}))} - ln\frac{P_{HN}(1-P_L)}{(P_L(1-P_{HN}))} \quad (8)$$

This method was applied to six pairs of codons for the amino acids: Phe, Ile, Tyr, Asn, Asp and His (FIYNHD) where the C-ending codon is always optimal, and for which the nature of selected codon usage bias appears to be the same (see Chapter 7). Here, this methodology is applied to the sequences of highly expressed genes but a modified version of the same method could be applied to lowly expressed genes, assuming the absence of efficiency-

selected codon usage bias. It was not possible to apply this version of the method to these Archaea due to the paucity of lowly expressed genes conserved across all species.

8.2.2 Data sources

The same 67 species of Archaea and their complete genome sequences, genome-wide codon usage and optimal growth temperatures were used as in Chapters 4 and 5. Genome-wide codon usage was assumed to reflect equilibrium frequencies in the absence of selection (P_L) since the fraction of highly expressed genes is a small subset of the total. A large dataset of homologous genes expected to be highly expressed and thus subject to selected codon usage bias was identified on the basis of genome annotation and evolutionary conservation. Genome annotation was used to identify 80 genes expected to be highly expressed (and thus subject to selected codon usage bias) on the basis of their functions as ribosomal protein genes, translation elongation factors, ATPases or RNA polymerases. BLAST (ALTSCHUL *et al.* 1990) was used to identify the orthologous copies of these genes in each of the 67 archaeal species. Genes without orthologues present in all of the 67 species were excluded, leaving 45 genes in the remaining dataset: *rpl1*, *rpl2*, *rpl3*, *rpl4*, *rpl5*, *rpl6*, *rpl7*, *rpl10*, *rpl10e*, *rpl11*, *rpl13*, *rpl14*, *rpl15*, *rpl15p*, *rpl18ae*, *rpl19*, *rpl21*, *rpl22*, *rpl32*, *rpl37*, *rpl44*, *rps2*, *rps3*, *rps4*, *rps5*, *rps6*, *rps7*, *rps8*, *rps8e*, *rps9*, *rps9p*, *rps10*, *rps11*, *rps12*, *rps13*, *rps15*, *rps17p*, *rps19*, *rps19p*, *tufA*, *fusA*, *atpA*, *atpB*, *rpoA*, *rpoB*. Values of S_{eff} and S_{acc} were calculated for each of the six amino acids and the final value was obtained as a weighted average of the numbers of codons.

8.2.3 Identifying conserved residues

Equilibrium frequencies of codon usage under the influence of accuracy- and efficiency-selected codon usage bias (P_{HC}) were estimated from conserved sites, whereas those frequencies under the influence of only efficiency-selected codon usage bias (P_{HN}) were estimated from non-conserved sites. Orthologous gene sequences for each of the 45 highly expressed genes were aligned, according to their protein sequence alignments using the Clustalw (THOMPSON *et al.* 1994) algorithm implemented in the BioEdit package (HALL 1999). Any ambiguous regions of the alignment and any sites containing one or more gaps were removed. Alignments were concatenated to produce a large alignment, 8857 codons in length, of which ~1800 encoded FIYNHD amino acids.

Alternative datasets of conserved sites were determined across varied evolutionary depths to examine their impact upon estimates of S_{acc}. In each case non-conserved sites were defined simply as the other remaining sites. First, the most highly conserved sites, common to all Archaea, were defined as those for which amino acids were invariant across the entire alignment. These accounted for ~10% of all sites. Second, the alignment was split into two datasets - each containing species from the two major archaeal kingdoms, the Crenarchaeota and Euryarchaeota, and four species which do not fall into either of these major kingdoms were excluded. Here, conserved sites were defined as those invariant within each of the two alignments, and accounted for near to 20% of all sites. Third, to consider conserved sites which include more recent local adaptation, the alignment containing all archaeal species was split into twelve separate alignments, each containing members of the same archaeal family. Conserved sites were then defined as those invariant within each family. Due to heterogeneity in divergence within each family, the resulting twelve datasets contained different numbers of conserved sites. To obtain conserved sites that were more easily comparable (encompassing around 55-65% of all sites), the most divergent members of each family were successively removed until around ~5500 of the total 8857 sites were defined as conserved. This final step resulted in the exclusion of three archaeal families: the Archaeoglobales, Methanosarcinales and Thermoplasmatales, where members exhibited too much or too little divergence to obtain comparable conserved sites. Thus three estimates of S_{acc} (for which there is some overlap in the designation of conserved sites) were obtained for each of 37 species of Archaea. Finally, values of S_{acc} were also computed for sites uniquely conserved within members of the Euryarchaeota. This allowed for comparison with values of S_{acc} computed across different independent sites conserved across all Archaea, to provide means of assessing the reliability of values.

8.2.4 Significance testing

To assess whether the strength of accuracy-selected codon usage bias (S_{acc}) was greater than zero, null distributions for each species were obtained. One thousand sets of randomly selected FIYNHD sites, corresponding to the observed number of conserved sites, were obtained from the alignment. Then values of S_{acc} based upon these sites were calculated for each species, and the 95% range of samples was recorded.

8.3 Results

8.3.1 The strength of accuracy-selected codon usage bias

The strength of accuracy-selected codon usage bias (S_{acc}) was estimated for each of 37 species of Archaea by comparing codon frequencies at conserved and non-conserved sites of highly expressed genes with genome-wide codon frequencies across six (FIYNHD) amino acids. Values of S_{acc} indicate the additional degree of selected codon bias at conserved relative to non-conserved sites of highly expressed genes. Three different estimates of S_{acc} were obtained for each species using sites conserved at varying evolutionary depths (Table 8-1). Estimates of the strength of efficiency-selected codon usage bias (S_{eff}) were co-estimated with S_{acc} values. The S_{eff} values reflect the intensity of selected codon usage bias across non-conserved sites of highly expressed genes. Across each of the three datasets, S_{eff} estimates were broadly similar, and were highly correlated with estimates of the total strength of selected codon usage bias (S) estimated in Chapter 5 (all $r > 0.94$, $p < 0.001$).

Species	Archaea (180)		Kingdom (320)		Family (1000)	
code	S_{acc}	Random	S_{acc}	Random	S_{acc}	Random
Methanomicrobiales						
Metmar	0.06	(-0.40/0.74)	0.21	(-0.34/0.45)	0.00*	(-0.87/-0.05)
Metboo	0.12	(-0.30/0.41)	0.19	(-0.23/0.34)	0.11	(-0.23/0.29)
Methun	0.45*	(-0.31/0.28)	0.21	(-0.24/0.21)	0.17*	(-0.32/0.10)
Metpal	0.19	(-0.32/0.38)	0.40*	(-0.28/0.25)	0.21*	(-0.63/-0.09)
Halobacteriales						
Halmar	0.84*	(-0.49/0.83)	0.18	(-0.38/0.69)	0.27	(-0.50/0.44)
Halsal	1.03*	(-0.73/0.79)	0.93*	(-0.63/0.58)	0.22	(-0.80/0.31)
Halwal	-0.21	(-0.33/0.24)	-0.07	(-0.23/0.22)	-0.21	(-0.12/0.26)
Natpha	2.00*	(-0.57/0.68)	1.66*	(-0.53/0.61)	0.22	(-0.65/0.34)
Hallac	0.90	(-0.49/0.92)	0.65	(-0.49/0.68)	0.01	(-0.70/0.40)
Haluta	0.60	(-0.47/0.80)	0.91*	(-0.39/0.69)	0.34*	(-0.83/0.12)
Halmuk	0.08	(-0.61/1.00)	0.42	(-0.48/0.95)	-0.14	(-0.60/0.67)
Haltur	0.70*	(-0.68/0.68)	1.12*	(-0.61/0.67)	0.42	(-0.45/0.66)
Natmag	0.84*	(-0.45/0.74)	0.67*	(-0.40/0.67)	0.55*	(-0.44/0.46)
Methanococcales						
Metjan	0.65*	(-0.31/0.32)	0.55*	(-0.26/0.25)	0.21*	(-0.32/0.20)
Metaeo	0.08	(-0.26/0.37)	0.06	(-0.15/0.32)	0.16	(-0.10/0.42)
MetmarS2	0.41	(-0.30/0.49)	0.42*	(-0.21/0.40)	0.40*	(-0.27/0.27)
Metvan	0.49*	(-0.30/0.34)	0.33*	(-0.21/0.29)	0.36*	(-0.14/0.32)
Metfer	0.79*	(-0.30/0.37)	0.64*	(-0.23/0.28)	0.29*	(-0.32/0.20)
Metvul	0.68*	(-0.35/0.26)	0.54*	(-0.27/0.24)	0.23*	(-0.36/0.16)
Methanobacteriales						
Metthe	1.12*	(-0.29/0.57)	1.00*	(-0.29/0.30)	0.48*	(-0.97/-0.13)
Metsmi	0.48*	(-0.32/0.31)	0.47*	(-0.19/0.26)	0.31*	(-0.28/0.14)
Metsta	0.80*	(-0.36/0.31)	0.54*	(-0.27/0.23)	0.18	(-0.26/0.22)
Metrum	0.76*	(-0.25/0.34)	0.70*	(-0.24/0.24)	0.31*	(-0.39/0.06)
Thermoproteales						
Calmaq	0.40*	(-0.45/0.32)	0.35*	(-0.35/0.22)	0.38*	(-0.46/0.20)
Pyraer	0.13	(-0.33/0.35)	0.18	(-0.22/0.25)	0.19*	(-0.46/0.08)
Pyrars	0.24	(-0.28/0.36)	0.37*	(-0.27/0.26)	0.47*	(-0.46/0.11)
Pyrcal	0.14	(-0.37/0.41)	0.33*	(-0.31/0.22)	0.41*	(-0.53/0.10)
Pyrisl	0.04	(-0.29/0.31)	0.08	(-0.23/0.22)	0.37*	(-0.28/0.21)
Theneu	0.07	(-0.39/0.58)	0.33	(-0.33/0.34)	0.47*	(-0.61/0.12)
Sulfolobales						
Metsed	0.19	(-0.32/0.25)	-0.01	(-0.26/0.18)	-0.03	(-0.35/0.18)
Sulaci	0.28*	(-0.38/0.27)	0.32*	(-0.28/0.19)	0.25*	(-0.44/0.20)
Sulsol	-0.04	(-0.44/0.24)	-0.06	(-0.31/0.18)	0.02	(-0.40/0.25)
Sultok	0.10	(-0.50/0.28)	0.10	(-0.29/0.25)	0.21	(-0.43/0.25)
Sulisl	0.02	(-0.42/0.23)	0.02	(-0.27/0.19)	-0.07	(-0.38/0.26)
Desulfurococcales						
Aerper	0.65*	(-0.33/0.54)	0.35	(-0.26/0.36)	0.20	(-0.57/0.27)
Hypbut	0.60*	(-0.35/0.39)	0.24*	(-0.31/0.21)	0.19*	(-0.63/-0.05)
Ignhos	1.30*	(-0.49/0.85)	0.59	(-0.47/0.64)	0.58*	(-1.37/-0.10)

Table 8-1 The strength of accuracy-selected codon usage bias (S_{acc}) across alternative conserved sites in Archaea

The strength of accuracy-selected codon usage bias (S_{acc}) is shown for 37 species of Archaea (codes in Appendix A). Values of S_{acc} are shown for sites conserved (i) among all Archaea (ii) within archaeal kingdoms (the Euryarchaeota and Crenarchaeota), and (iii) within archaeal families (shown). Numbers in heading brackets indicate the approximate number of conserved sites for which estimates of S_{acc} are based upon. Asterisks indicate significant values, greater than the upper bounds of the 95% range of S_{acc} values across 1000 sets of randomly selected conserved sites (random).

Estimates of S_{acc} based upon individual amino acid frequencies were correlated e.g. (Figure 8-1) but not as highly as components of S_{eff}. Similarly, a positive correlation was observed with estimates of S_{acc} based upon sites conserved among all Archaea (n~180) with those values based upon sites uniquely conserved across the Euryarchaeota (n~140) albeit noisily ($r = 0.47$, $p < 0.01$).

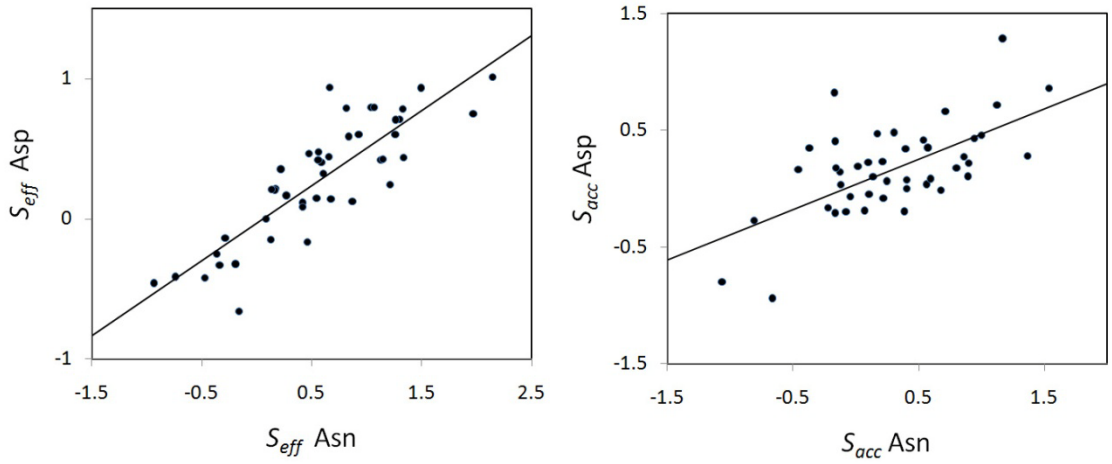


Figure 8-1 The strengths of efficiency- (S_{eff}) and accuracy-selected codon usage bias (S_{acc}) across amino acid components Asn and Asp in Archaea

Values shown for sites conserved within familes (n ~1000) for 37 species of Archaea. Components Asn and Asp are correlated for S_{eff} ($r = 0.85$, $p < 0.001$) and S_{acc} ($r = 0.63$, $p < 0.001$) values.

It has been suggested that positive associations of optimal codons with conserved sites (and thus S_{acc} values) might arise if a fraction of non-conserved sites reflect non-synonymous mutations from optimal to non-optimal codons (AKASHI 1994). Of the non-optimal codons among the FIYNHD sites examined here (which are all U-ending), only one single base pair

substitution from the commonly optimal codon GGU for glycine (Chapter 7) to the non-optimal codon GAU for aspartic acid could give rise to these patterns. Values of S_{acc} for aspartic acid were no higher than values for other amino acids with the same average value (S_{acc} = 0.18), suggesting that mutations from optimal codons were unlikely to be a major factor determining the S_{acc} values obtained here.

The vast majority (92%) of S_{acc} values are positive (Table 8-1) with 95% of values ranging from -0.34 to 1.11, indicating additional selected codon bias at conserved relative to non-conserved sites. For approximately half of the species in each of the three datasets, values of S_{acc} were greater in magnitude than the upper 95% range of values from the null distribution and so are conventionally statistically significant. However there appear to be some reliability issues since the species for which S_{acc} values are significant vary according the dataset of conserved sites; S_{acc} values are significant in one of three datasets for seven archaeal species, and in two of the three datasets for 10 species. This seems poor considering that there is some overlap in the conserved sites within each dataset; all datasets contain the ~180 residues conserved all Archaea. Nevertheless there are 11 species for which S_{acc} values are significant across all three datasets. Six of these species are from the sister taxa the Methanococcales and the Methanomicrobiales, indicating that S_{acc} values exhibit some phylogenetic dependence. Two of the species with consistently significant S_{acc} values, *Sulfolobus acidocaldarius* and *Caldivirga maquilingensis*, exhibit low (indeed negative) efficiency-selected codon usage bias (S_{eff} = -0.38 and -0.27 respectively) and were not found to exhibit significant *S* values in Chapter 5.

8.3.2 Variation in the strength of accuracy-selected codon usage bias

Estimates of S_{acc} vary according to the dataset of conserved sites that is used. Typically the highest values are obtained for species across the dataset containing the most highly conserved sites (Figures 8-2 and 8-3). As larger numbers sites are included, conserved across more recent evolutionary time scales, values of S_{acc} are reduced in magnitude, and distributions of S_{acc} values across each of the three datasets differ significantly (paired Wilcoxon tests, $p < 0.001$). There are however many exceptions to this general trend. For instance, values for *Methanococcus maripaludis* remain relatively unchanged across all three datasets (Figure 8-2).

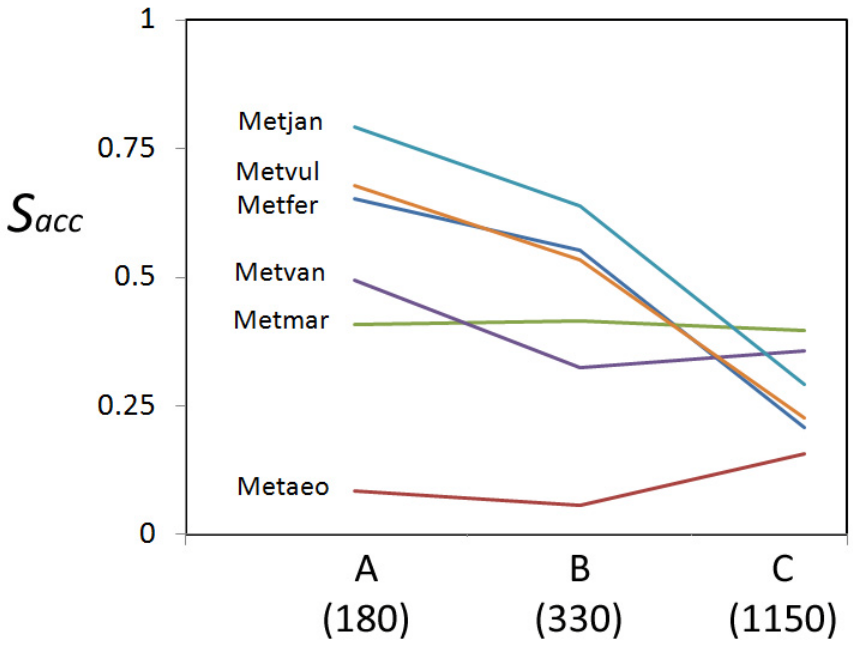


Figure 8-2 The strength of accuracy-selected codon usage bias (S_{acc}) across alternative conserved sites for species of the Methanococcales

Shown for values of S_{acc} based upon sites conserved (A) among all Archaea (B), within archaeal kingdoms and (C) within archaeal families. Numbers in brackets indicate the approximate numbers of conserved sites from a total of ~1800 sites. Lines connect values for the same species.

Values for the strength of efficiency-selected codon usage bias (S_{eff}) do not vary substantially according to the designation of non-conserved sites (Figure 8-3), and across all datasets of conserved sites, S_{eff} values are consistently greater than S_{acc} values (paired Wilcoxon tests, $p < 0.001$). The dataset of sites conserved within archaeal families (n ~ 1000; 55%-65% of all sites conserved) was selected for subsequent analyses as this is expected to exhibit the lowest degree of sampling error of all datasets examined.

The reliability of S_{acc} values is likely to depend greatly upon the reasonability of the assumption that it is all and only the most highly conserved sites which are of greatest structural and functional importance. An argument can be made that, given the rapid rates of adaptive evolution estimated for prokaryotes (CHARLESWORTH and EYRE-WALKER 2006), non-conserved sites with adaptive substitutions are likely to be equally as important as those which have been subject to selective constraint for millions of years. Here, only the most

highly conserved genes for which structure and function remain largely unchanged have been analysed. The results here indicate that non-conserved (potentially adaptive) sites among these genes are in fact depleated in optimal codons, since S_{acc} values are highest when they are estimated based upon the smallest subset of the most highly conserved sites. Thus it seems that in this case, the assumption that the most highly conserved sites are the most important is reasonable.

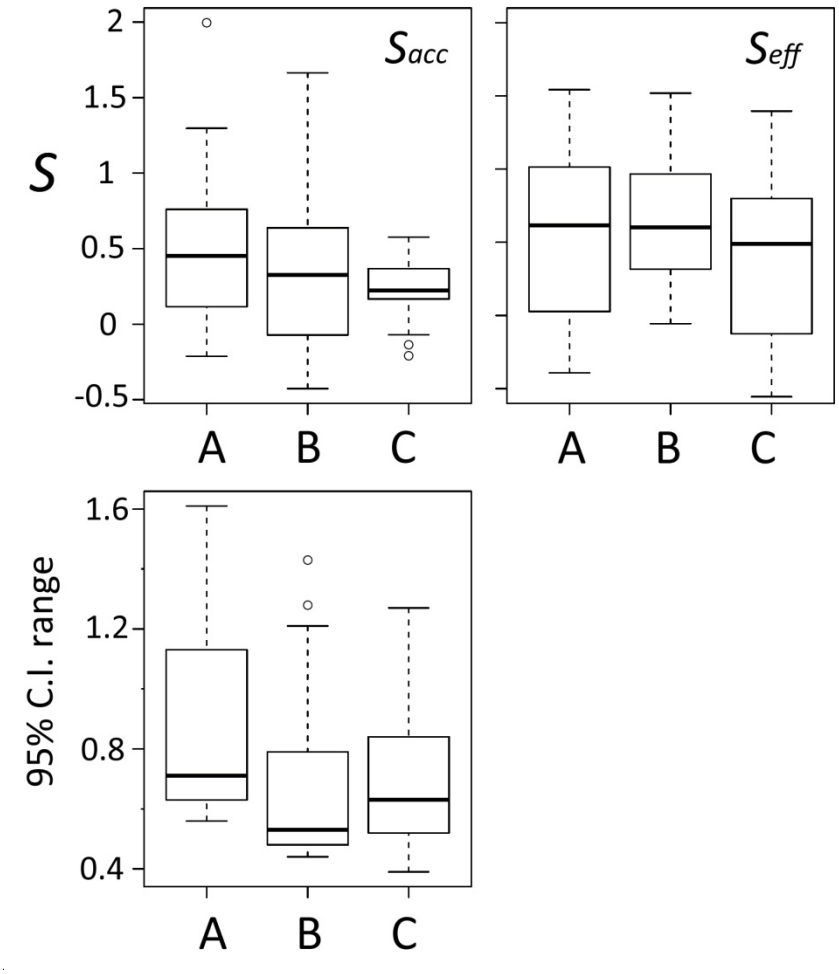


Figure 8-3 Distributions of the strength of accuracy- (S_{acc}) and efficiency- (S_{eff}) selected codon usage bias and range of S_{acc} confidence intervals (C.I.) across alternative conserved sites for 37 species of Archaea

Values of S_{acc} and S_{eff} are based upon sites conserved (A) among all Archaea (B) within archaeal kingdoms and (C) within archaeal families. Dark horizontal lines indicate median values; boxed areas indicate the 50% range of values; error bars indicate the 95% range of values. Circles indicate values beyond the 95% range.

Phylogenetic independent contrasts were used to explore the relationship of S_{acc} and S_{eff} values among species, and no correlation was observed ($r = 0.06$; $p = 0.35$). To examine the impact of optimal growth temperature upon values of S_{eff} and S_{acc}, species were binned into two categories according to optimal growth temperature (above and below 40°C; Figure 8-4). Independent contrasts of S_{acc} and S_{eff} were found to be correlated within each temperature class ($r = 0.42$, $p = 0.06$ and $r = 0.65$, $p < 0.01$ respectively). The distributions of efficiency-selected codon usage bias were found to differ significantly between these categories, with lower values among thermophilic species (t-test, $p < 0.001$), but the distributions of accuracy-selected codon usage bias did not differ significantly among groups (Wilcoxon test, $p = 0.24$). Using other temperature thresholds or conservation criteria did not qualitatively affect the result.

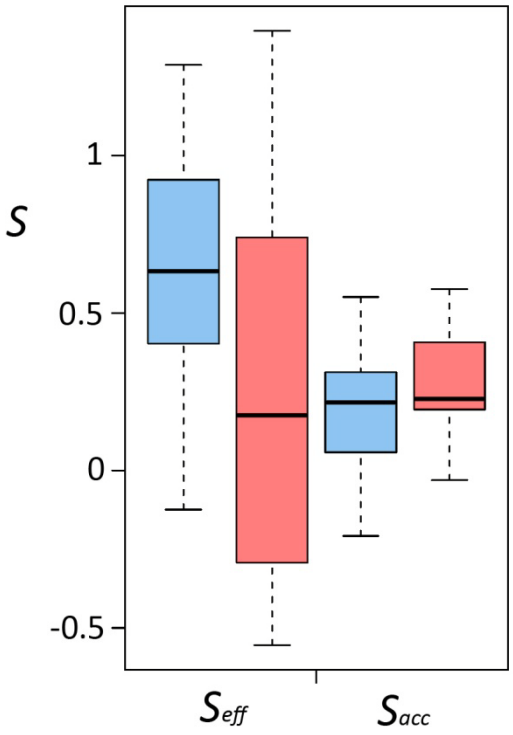


Figure 8-4 Distributions of the strengths of accuracy- (S_{acc}) and efficiency-selected (S_{eff}) codon usage bias in species of Archaea with different optimal growth temperatures

Coloured bars indicate the distribution of values of S_{acc} and S_{eff} across the dataset of sites conserved within archaeal families (n ~1000) for species with optimal growth temperatures (i) < 40°C (blue) and (ii) > 40°C (red). Lines as in Figure 8-3.

8.4 Discussion

8.4.1 Accuracy-selected codon usage bias?

This preliminary analysis has explored the independent estimation of accuracy- (S_{acc}) and efficiency-selected codon usage bias (S_{eff}). To disentangle these effects, the strength of accuracy-selected codon usage bias was assumed here to reflect additional selected codon usage bias across conserved relative to non-conserved sites. Patterns of selected codon usage bias observed in Chapter 5 indicated that this was a reasonable assumption to make. If accuracy-selected codon usage bias is stronger than expected, and impacts substantially upon non-conserved sites, then S_{acc} values are underestimates but nevertheless the relative magnitudes of S_{acc} values among species are unlikely to be affected.

Previous analyses have interpreted associations of optimal codons with conserved sites as reflecting accuracy-selected codon usage bias (AKASHI 1994; DRUMMOND and WILKE 2008; STOLETZKI and EYRE-WALKER 2007). There is some confidence that estimates of S_{acc} indeed reflects accuracy-selected codon usage bias since there is some constistency with (i) independent estimates based upon different amino acid residues and (ii) independent estimates based upon different sets of conserved sites. The weaker correlations observed among amino acid components of S_{acc} compared with those of S_{eff} are expected since values of S_{acc} are based upon three sources of codon frequency data (conserved, non-conserved and genome-wide) rather than two (non-conserved and genome-wide), and thus subject to greater sampling error. Additionally, the observation that estimates of S_{acc} were no higher for aspartic acid than for other amino acids, when its non-optimal synonym could arise at a non-conserved site by a single mutation from the optimal codon for glycine, indicates that these S_{acc} values are unlikely to reflect recent patterns of mutation.

Thus S_{acc} values appear to reflect accuracy-selected codon usage bias. Since the vast majority of values are positive, with around half of values significantly greater than expected, this appears to indicate that accuracy-selected codon usage bias is a feature of many Archaea. Whilst there has been some suggestion in the early biochemical literature that the most efficient codons are the least accurate (DIX and THOMPSON 1989; PARKER and PRECUP 1986),

the occurrence of predominantly positive S_{acc} values suggests that at least among FIYNHD codons, the identities of efficiency- and accuracy-selected optimal codons are the same.

8.4.2 The impact of dataset selection upon estimates of accuracy-selected codon usage bias

Several analyses have investigated the presence of accuracy-selected codon usage bias in the bacterium *Escherichia coli* by examining the presence of an association of optimal codons with conserved sites. First, in analyses of 118 genes in *E. coli* (HARTL *et al.* 1994), no enrichment of optimal codons was detected across sites conserved with the related species *Salmonella enterica* throughout 100 million years of evolution (OCHMAN and WILSON 1987). These Bacteria share ~96% of amino acid sequence identity (SHARP 1991), and so only about 4% of sites were assigned as non-conserved. Second, a similar approach was taken by a more recent study which was able to identify 2,229 orthologous genes from *E. coli* and *S. enterica* complete genome sequences (DRUMMOND and WILKE 2008). Across a dataset of all 2,229 orthologs, the frequency of optimal codons was found to be significantly higher across conserved sites. Yet across a restricted dataset of the 10% most highly expressed genes, where accuracy-selected codon usage bias is expected to be most important (WILKE and DRUMMOND 2006), the elevated frequency of optimal codons across conserved sites was not found to be significant. It was speculated that this was due to the occurrence of very few non-optimal codons among highly expressed genes (DRUMMOND and WILKE 2008). However this seems unlikely as the highest F_{op} values in *E. coli* are < 0.9, indicating that more than 10% of sites (even across the most highly expressed genes) are non-optimal, and so could easily encompass the 4% of non-conserved sites analysed. Alternatively, it has been suggested that the high fraction of adaptive substitutions estimated in *E. coli* might mean that selective constraint is a poor indicator of the functional and/or structural importance of residues (STOLETZKI and EYRE-WALKER 2007). However, there was no indication that the rate of adaptive evolution varies with expression level (CHARLESWORTH and EYRE-WALKER 2006) and so is unlikely to explain the discrepancy in the Drummond and Wilke study. Nevertheless, when the definition of conserved sites was altered to encompass all sites besides a small fraction (<1%) with amino acid polymorphisms, optimal codons were found to be significantly enriched across conserved sites (STOLETZKI and EYRE-WALKER 2007).

Collectively these studies indicate that the detection of accuracy-selected codon usage bias for a single species is somewhat sensitive to the dataset of conserved sites.

This study examined the impact of dataset selection upon the strength of accuracy-selected codon usage bias using datasets of sites conserved across three different evolutionary depths. Even the shallowest evolutionary depth used in this study included sites conserved throughout an estimated billion years of evolution (BATTISTUZZI *et al.* 2004) - an order of magnitude greater than *E. coli* and *S. enterica* divergence, and thus including a more even ratio of conserved to non-conserved sites. A trend emerged whereby the highest values of S_{acc} were obtained across the most highly conserved sites, consistent with selective constraint as a reliable indicator of functional and/or structural importance, and indicating that accuracy is most relevant to a minority of the most important sites. If this pattern is indicative of a trend among all species, it may explain why previous studies which have considered sites conserved over shorter evolutionary time periods have only detected very small effects (AKASHI 1994; DRUMMOND and WILKE 2008; STOLETZKI and EYRE-WALKER 2007). However, not all of the species examined conformed to this trend and many inconsistencies were observed in the relative magnitudes of S_{acc} values among species. Some reliability issues are expected due to sampling error, which is expected to be greatest across estimates based upon the fewest most highly conserved sites. It seems that not all of these inconsistencies can be explained by sampling error since in most cases, the range of null values was lowest across the intermediate evolutionary depth of sites conserved within archaeal kingdoms. Further work is required to verify the repeatability of these estimates.

8.4.3 Trends and future directions

This study has provided the first estimates of the strength of accuracy-selected codon usage bias. Estimates were obtained for each of 37 species of Archaea to examine patterns of variation which might shed light on some evolutionary questions. One question of interest was how the strength of accuracy-selected codon usage bias varies with the strength of efficiency-selected codon usage bias. When all values of S_{acc} and S_{eff} were considered, no correlation of these indices was observed, however, a positive correlation was revealed when species were binned into two cateogries according to optimal growth temperature. The positive correlation implies that either accuracy-selection conforms to an 'accuracy for

efficiency' model of codon selection, or there is indeed no efficiency-selection at all, and all selection is for accuracy. The latter explanation seems unlikely to be the case since the indices of S_{eff} and S_{acc} behave different according to optimal growth temperature, indicating that they do not reflect the same underlying process. Whilst estimates of the strength of efficiency-selected codon usage bias were found to be reduced at high growth temperatures, consistent with the kinetic impact of growth temperature upon translation efficiency (Chapter 5); estimates of accuracy-selected codon usage bias were not. At high growth temperatures, errors in translation are expected to be more costly, and so the selective benefit of accuracy-selection is expected to increase (Chapter 5). If however, accuracy is for cellular efficiency, then it could be that the increase in accuracy-selection expected at high temperatures is offset by the reduced level of efficiency-selection, such that the overall effect is for accuracy-selection to remain unchanged at different growth temperatures (Figure 8-4).

Across all datasets examined, estimates of S_{eff} were consistently greater than those for S_{acc}. However, these analyses are unable to draw any firm conclusions as to the relative importance of efficiency- and accuracy-selected codon usage bias since they assume that accuracy-selection only influences non-conserved sites. Even so, across all datasets examined, it is not clear that if this assumption was relaxed, (i.e. by including a dataset with a much larger fraction of conserved sites), that this would substantially impact upon estimates of S_{acc} because these estimates were highest for the dataset with least number of the most highly conserved sites. It will be interesting to see if this finding is validated when the same approach is applied to bacterial genomes, where the wide range of the overall S values allows for clearer examination of trends. Validation could come from a number of sources including obtaining further datasets of orthologous genes of varying expression levels. It will be interesting to extend these analyses to the conserved and non-conserved sites of lowly expressed genes where efficiency-selected codon bias is expected to be absent to better address this question.

9. CONCLUSIONS

9.1 Codon usage bias in the third domain of life

There were reasons to suspect *a priori* that biases in synonymous codon usage in Archaea would differ from those observed in the well-studied bacterial domain. The domains differ in aspects of biology and ecology which are expected to be relevant to codon bias. Archaea contain fewer rRNA operons and tRNA genes (HENRY 2007), and a substantially diverged translation machinery (LECOMPTE *et al.* 2002), and are specialists in a variety of unusual habitats. This thesis has characterised patterns of codon usage bias in the Archaea, the third and understudied domain of life.

The analyses commenced with the investigation of patterns of codon usage bias in the well-studied and characteristic archaeon, *Methanococcus maripaludis*. By comparison with the archetypal bacterium *Escherichia coli*, the archaeon exhibited a high intensity of selected codon usage bias given its minimal set of tRNA genes and moderate growth rate. Yet unlike the bacterium, selection was largely restricted to two-fold degenerate sites. To explore whether codon usage bias in *M. maripaludis* was typical of Archaea, various analyses were used to identify patterns of synonymous codon usage in a further 66 species for which complete genome sequences were available. Multivariate analyses were used to examine trends in codon usage bias both within, and among, the genomes of Archaea. From these it seemed that the dominance of translational selection upon patterns of codon usage in *M. maripaludis* was atypical. Of the major trends in codon usage bias identified within archaeal genomes, few, by comparison with an analogous study of bacterial genomes (SUZUKI *et al.* 2008), were associated with expression level and interpreted as reflecting translational selection. Distributions of additional intragenomic trends associated with heterogeneity in GC3s and strand-skew indices did not differ substantially between the domains. Nonetheless systematic patterns of GC3s variation were observed across some archaeal chromosomes and these merit further investigation. Similarly, in analyses among genomes, the same two major trends which explain most of the variation in codon usage between bacterial genomes (CHEN *et al.* 2004; LYNN *et al.* 2002), were also apparent across Archaea. Two additional trends of minor effect were identified describing variation in codon bias among archaeal genomes, and it remains to be seen whether these are also apparent among

the genomes of Bacteria. Thus from these analyses, the major difference in patterns of codon bias among Archaea and Bacteria appeared to be in the lack of prevalence of translational selection.

To investigate this apparent difference, the strength of selected codon usage bias (S) was estimated for each of the 67 archaeal species. Consistent with the low numbers of rRNA operons and tRNA genes observed among Archaea, there was an absence of very high values of S observed for some bacterial genomes (SHARP *et al.* 2005). Unlike bacterial genomes, variation in the strength of selected codon usage bias was not found to be influenced by rRNA operon or tRNA gene numbers. Rather, a linear model revealed that only minimal generation time and growth temperature explained significant variation in S. When values of S were binned according to optimal growth temperature, they were found to be negatively correlated with generation time, with values of S systematically lower in species living at high temperatures. Since the distribution of minimal generation times across Archaea was not different from those of Bacteria, it seems that the generally lower magnitude of S values observed in Archaea may be largely attributed to the relatively high fraction of thermophilic species. What is less clear, when adaptation to new growth temperatures can occur over relatively short evolutionary timescales (PUIGBO *et al.* 2008), is why the distributions of thermophilic species differ between Archaea and Bacteria.

9.2 Kinetic impact of growth temperature upon selected codon usage bias

There are three possible explanations for why values of S are systematically lower in species which live at high temperatures. The first is if there is a systematic reduction in the effective population size (N_e) with temperature, meaning that estimates of S (expected to correspond to $2N_eS$) are lower in thermophilic species. There is no evidence that this is the case, and it seems unlikely since the effects of N_e are expected to reduce the effectiveness of all forms of selection and yet purifying selection upon protein sequences appears to be strongest in thermophilic species (FRIEDMAN *et al.* 2004), and here estimates of the strength of accuracy-selected codon usage bias were no weaker in thermophilic species. The second possible explanation is that species living at high temperatures engage in less competitive lifestyles, perhaps due to reduced levels of interspecific competition associated with specialisation. If

this were the case then thermophilic species are expected to exhibit slower maximal growth rates, and yet here, faster growth rates were observed in species living at high temperatures.

Finally, *S* could be reduced among thermophilic species due to the kinetic impact of growth temperature upon translation elongation. Kinetic theory predicts that increases in temperature increase the rate of translation elongation. Therefore rapid elongation rates at high temperatures mean that translation is intrinsically more efficient, thus reducing the selective benefit of optimal codon usage for the efficiency of translation. Provided that elongation rate is to some extent growth-limiting, as has been demonstrated experimentally in *E. coli* (FAREWELL and NEIDHARDT 1998), then the kinetic impact of growth temperature on elongation rate could explain why species living at high temperatures grow rapidly. This has potential implications for life history strategies, leading to the prediction that species living at high temperatures invest fewer resources in their translation machinery. Consistent with this, the numbers of rRNA operons and tRNA genes, taken as a proxy for translational investment, were found to be negatively correlated with optimal growth temperature. Thus it seems that the kinetic impact of growth temperature upon elongation rate is the most likely explanation why *S* values are reduced with increasing growth temperatures.

9.3 The nature of selected codon usage bias

While it seems that selection for growth efficiency is less relevant to species living at high temperatures, there is a growing body of literature which suggests that selection for fidelity of many biological processes may be increased. Fidelity of DNA replication is greater in thermophilic species where mutation rates are reduced (DRAKE 2009). Fidelity of protein folding is expected to be more relevant at high temperatures (BLOOM *et al.* 2005), consistent with elevated abundances of chaperone proteins (PHIPPS *et al.* 1993; VAN BOGELEN *et al.* 1992). Fidelity of protein sequence appears to be more critical in thermophilic species, where proteins appear to evolve more slowly (FRIEDMAN *et al.* 2004) and are of similar amino acid composition to those of highly expressed genes (CHERRY 2010). Finally, the fidelity of metabolic networks also appears to be more important in thermophilic species, where network structures are more robust and less efficient (TAKEMOTO and AKUTSU 2008). Collectively these observations seem to indicate that accuracy of a various biological processes is favoured over efficiency at higher growth temperatures.

It has been suggested that transltional accuracy is the primary target of translational selection in a wide range of organisms (DRUMMOND and WILKE 2008). If this is the case, then the importance of fidelity at high temperatures is expected to extend to selected codon usage bias because the fitness costs of translational missense errors are likely to reflect those of mutations to gene sequences (WILKE and DRUMMOND 2006), which are more costly at high temperatures (DRAKE 2009; LYNCH 2010). Therefore accuracy-selected codon usage bias to avoid errors in translation is likely to be most important at high temperatures where obtaining the correct protein sequence is most critical (BLOOM *et al.* 2005; FRIEDMAN *et al.* 2004). However, these analyses have revealed that the total selected codon usage bias was reduced at high temperatures in Archaea, and may only be explained if the major selective benefit of optimal codon usage is in the efficiency of translation. In this situation, the benefit of selected codon usage bias is reduced at high temperatures since elongation rates are increased. Thus whilst selection to avoid protein misfolding may be a dominant constraint on protein evolution (DRUMMOND 2006), it was less clear whether it has any role in shaping the evolution of synonymous sites in Archaea.

The potential effects of accuracy-selected codon usage bias were explored. Previous analyses have interpreted alternative patterns of codon usage across conserved and non-conserved sites as evidence for accuracy-selected codon usage bias (AKASHI 1994; STOLETZKI and EYRE-WALKER 2007). Following on from these analyses, here a method was developed to obtain independent estimates of the relative strengths of efficiency- and accuracy-selected codon usage bias. Initial estimations suggested that there was indeed evidence for accuracy-selected codon usage bias across the highly expressed gene sequences of some species of Archaea. Accuracy-selection was found to be largely restricted to a small subset of the most highly conserved sites, consistent with its relatively minor role by comparison with efficiency-selected codon usage bias.

9.4 Optimal codon divergence

Initial analyses of the codon usage in *M. maripaludis* identified that, unlike the bacterium *E. coli*, selection was reduced at four-fold relative to two-fold degenerate sites. To determine whether this was a widespread phenomenon among Archaea, the strength of selected codon usage bias was estimated across two-fold (*S*2) and four-fold (*S*4) degenerate sites, and in all species examined, *S*2 was greater than *S*4. One possible explanation for this relates to the different tRNA requirements of each degeneracy class. Four-fold degenerate sites are decoded by larger numbers of different tRNA anticodons than two-fold degenerate sites. If each of the different tRNA anticodons is best at translating a different synonymous codon then conflicting selective effects may arise whereby different forms of tRNA favour different optimal codons, resulting in a reduction in the selective benefit of the overall optimal codon. This potential effect is expected to impact most greatly upon four-fold and six-fold degenerate sites where there are the largest numbers of tRNA anticodons. This hypothesis was tested by comparing selection across sites decoded by either one or two different tRNA anticodons. In direct conflict with predictions, selection was found to be greater across sites decoded by two rather than one form of tRNA.

An alternative explanation relates to how the identities of optimal codons evolve to vary among species. Optimal codon divergence was investigated in Archaea and was found to be largely restricted to four and six-fold degenerate amino acid groups, with optimal codons invariant for U and C-ending two-fold degenerate groups. The identity of optimal codons are expected to diverge in response to directional changes in mutational bias (SHIELDS 1990) and patterns of codon usage bias accross Bacteria are consistent with this (HENRY 2007). Here, among Archaea, both the G+C content across third optimal codon positions, as well as G+C content across first tRNA anticodon positions, were found to vary with genome-wide GC3s across four and six-fold degenerate sites, consistent with a major role for mutational biases in optimal codon divergence. What is less clear is whether mutational biases drive switches in codon usage bias in the presence, or absence, of translational selection (SHIELDS 1990). Optimal codon divergence following the relaxation of selected codon usage bias has the potential to explain why selection across four-fold degenerate sites (*S*4), where optimal codons are observed to diverge, is weaker than selection across two-fold degenerate sites (*S*2), where optimal codons are largely invariant. The impact of optimal codon divergence

upon estimates of *S* was investigated. Optimal codon divergence events were mapped to the phylogeny of Archaea, and values of *S* corresponding to these optimal codons were estimated in species prior to and subsequent to optimal codon divergence. As expected if optimal codon divergence occurs following the relaxation of selected codon bias, values of *S*, reflecting the long-term effectiveness of selected codon usage bias, were lower across species where optimal codons had been observed to diverge. So it seems that these analyses indicate that optimal codon divergence proceeds via the relaxation of translational selection, and may explain why values of *S*4 are lower than *S*2 in Archaea. Yet several issues remain. How and why does the relaxation of selected codon usage bias for some amino acids, but not others, occur? Is optimal codon divergence ever driven solely by directional mutational pressure in spite of natural selection? Do these observations, based upon limited numbers of optimal divergence events, reflect general trends in the evolution of selected codon usage bias? It will be interesting to extend these analyses to the bacterial domain to explore these questions.

LITERATURE CITED

AKASHI, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927-935.

AKASHI, H., R. M. KLIMAN and A. EYRE-WALKER, 1998 Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. Genetics **102/103:** 49–60.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403-410.

ALVAREZ, F., C. ROBELLO and M. VIGNALI, 1994 Evolution of codon usage and base contents in kinetoplastid protozoan. Mol. Biol. Evol. **11:** 790–802.

ANDERSSON, G. E., and P. M. SHARP, 1996 Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology **142:** 915-925.

ANDERSSON, S. G. E., and C. G. KURLAND, 1990 Codon preferences in free-living microorganisms. Microbiol. Rev. **54:** 198-210.

ARGOS, P., M. G. ROSSMAN, U. M. GRAU, H. ZUBER, G. FRANK *et al.*, 1979 Thermal stability and protiein structure. Biochemistry **18:** 5698-5703.

BASTOLLA, U., A. MOYA, E. VIGUERA and R. C. H. J. HAM, 2004 Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J. Mol. Biol. **343:** 1451-1466.

BATTISTUZZI, F. U., A. FEIJAO and S. B. HEDGES, 2004 A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colinization of land. BMC Evolutionary Biology **4:** 44.

BAUMANN, P., L. BAUMANN and M. A. CLARK, 1996 Levels of *Buchnera aphidicola* chaperonin GroEL during growth of the aphid *Schizaphis graminum*. Curr. Microbiol. **32:** 279-285.

BENNETZEN, J., and B. HALL, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

BERG, O. G., and C. G. KURLAND, 1996 Growth rate-optimised tRNA abundance and codon usage. J. Mol. Biol. **270:** 544-550.

BERG, O. G., and M. MARTELIUS, 1995 Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. J. Mol. Evol. **41:** 1432-1432.

BERNARDI, G., and G. BERNARDI, 1986 Compositional constraints and genome evolution. J. Mol. Evol. **24:** 1-11.

BERNARDI, G., B. OLOFSSON, I. J. FILIPSK, M. ZERIAL, SALINAS. J. *et al.*, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953-958.

BERQUIST, B. R., and S. DAS SARMA, 2003 An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, *Halobacterium sp*. strain NRC-1. J. Bacteriol. **185:** 5959-5966.

BIRDSELL, J. A., 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19:** 1181-1197.

BLOOM, J. D., J. J. SILBERG, C. O. WILKE, D. A. DRUMMOND, C. ADAMI *et al.*, 2005 Thermodynamic prediction of protein neutrality. Proc. Nat. Acad. Sci. USA **102:** 606-611.

BROWN, T. C., and J. JIRICNY, 1989 Repair of base-base mismatches in simian and human cells. Genome Biol. **31:** 578-583.

BULMER, 1991a Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728-730.

BULMER, M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728-730

BULMER, M., 1990 The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res. **18:** 2869-2873.

BULMER, M., 1991b The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897-907.

CAMPBELL, W. H., and G. GOWRI, 1990 Codon usage in higher plants, green algae, and cyanobacteria. Plant Physiol. **92:** 1-11.

CHAMARY, J. V., J. L. PARMLEY and L. D. HURST, 2006 Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. **7**: 98-108.

CHARIF, D., J. THIOULOUSE, J. R. LOBRU and G. PERRIERE, 2005 Online synonymous codon usage analyses with the ade4 and seqinR packages Bioinformatics **21:** 545-547.

CHARLESWORTH, J., and A. EYRE-WALKER, 2006 The rate of adaptive evolution in Enteric Bacteria Mol. Biol. Evol. **23:** 1348-1356.

CHEN, S. L., W. LEE, A. K. HOTTES, L. SHAPRIO and H. H. MCADAMS, 2004 Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA **106:** 3480-3485

CHERRY, J. L., 2010 Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. Mol. Biol. Evol. **27:** 735-741.

CONDON, C., D. LIVERIS, C. SQUIRES, I. SCHWARTZ and C. L. SQUIRES, 1995 rRNA operon multiplicity in *Escherichia coli* and the physiological implications of rrn inactivation. J. Bacteriol. **177:** 4152-4156.

CRICK, F. H., 1966 Codon-anticodon pairing: the wobble hypothesis. J. Mol. Biol. **19:** 548-555.

CURRAN, J., and M. YARUS, 1989 Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J. Mol. Biol. **209:** 65–77.

CUTTER, A. D., 2008 Multilocus patterns of polymorphism and selection across the X chromosome of *Caenorhabditis remanei*. Genetics **178:** 1661-1672.

CUTTER, A. D., S. E. BAIRD and D. C. CHARLESWORTH, 2006 High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. Genetics **174:** 901-912.

DALE, C., B. WANG, N. MORAN and H. OCHMAN, 2003 Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol. Biol. Evol. **20:** 1188-1194.

DAS, S., S. PAUL, S. K. BAG and C. DUTTA, 2006 Analysis of *Nanoarchaeum equitans* genome and proteome composition: indications for hypothermophilic and parasitic adaptation. BMC Genomics **7:** 186.

DAUBIN, V., E. LERAT and G. PERRIÈRE, 2003 The source of laterally transferred genes in bacterial genomes. Genome Biol. **4:** R57.

DAUBIN, V., and G. PERRIERE, 2003 G+C3 Structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. **20:** 471-483.

DETHLEFSEN, L., and T. M. SCHMIDT, 2005 Differences in codon bias cannot explain differences in translational power among microbes. BMC Bioinformatics **6:** 3.

DETHLEFSEN, L., and T. M. SCHMIDT, 2007 Performance of the translational apparatus varies with the ecological strategies of Bacteria. J. Bacteriol. **189:** 3237-3245.

DIX, D. B., and R. C. THOMPSON, 1989 Codon choice and gene expression: synonymous codons differ in translational accuracy. Proc. Natl. Acad. Sci. USA **86:** 6888-6892.

DOBRINDT, U., B. HOCHHUT, U. HENTSCHEL and J. HACKER, 2004 Genomic islands in pathogenic and environmental microorganisms. Nat. Rev. Microbiol. **2:** 414-424.

DONG, H., L. NILSSON and C. G. KURLAND, 1996 Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. **260:** 649-663.

DOS REIS, M., R. SAVVA and L. WERNISCH, 2004 Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32:** 5036-5044.

DOS REIS, M., and L. WERNISCH, 2008 Estimating translational selection in eukaryotes. Mol. Biol. Evol. **26:** 451-461.

DRAKE, J. W., 1991 A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. USA **88:** 7160-7164.

DRAKE, J. W., 2009 Avoiding dangerous missense: thermophiles display especially low mutation rates. PLoS Genetics **5:** e1000520.

DRESSAIRE, C., C. GITTON and P. LOUBIÈRE, 2009 Transcriptome and proteome exploration to model translation efficiency and protein stability in *Lactococcus lactis*. PLoS Comput. Biol. **5:** e1000606.

DRUMMOND, D. A., 2006 Misfolding dominates protein evolution. Dissertation (Ph.D.), California Institute of Technology.

DRUMMOND, D. A., and C. O. WILKE, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134:** 341-352.

DURET, L., 2000 tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed. Trends in Genetics **16:** 287–289.

DURET, L., and P. F. ARNDT, 2008 The impact of recombination on nuceotide substitution in the human genome. PloS Genet. **4:** e1000071.

DURET, L., and N. GALTIER, 2009 Biased gene conversion and the evolution of mammalian genomic landscapes. Ann. Rev. Genom. & Hum. Genet. **10:** 285-311.

EHRENBERG, M., and C. G. KURLAND, 1984 Costs of accuracy determined by a maximal growth rate constraint. Quart. Rev. Biophys. **17:** 45-82.

EYRE-WALKER, A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 675-683.

EYRE-WALKER, A., and M. BULMER, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21:** 4599-4603.

EYRE-WALKER, A., and M. BULMER, 1995 Synonymous substitution rates in Enterobacteria. Genetics **140:** 1407-1412.

EYRE-WALKER, A., and L. D. HURST, 2001 The evolution of isochores. Nat. Rev. Genetic. **2**: 549-555.

FAREWELL, A., and F. C. NEIDHARDT, 1998 Effect of temperature on in vivo protein synthetic capacity in *Escherichia coli* J. Bacteriol. **180:** 4704-4710.

FELSENSTEIN, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1-15.

FILIPSKI, J., J. P. THIERY and G. BERNARDI, 1973 An analysis of the bovine genome by Cs2SO4-Ag+ density centrifugation. J. Mol. Biol. **80:** 177-197.

FOX, G. E., STACKEBRANDT, E., HESPELL, R.B., GIBSON, J., MANILOFF, J., DYER, T.A., WOLFE, RS., BALCH, W.E., TANNER, R.S., MAGRUM, LJ., ZABLEN, L.B., BLAKEMORE, R., GUPTA, R., BONEN, L., LEWIS, B.J., STAHL, D.A., LUEHRSEN, K.R., CHEN, K.N. AND WOESE, C.R., 1980 The phylogeny of prokaryotes. Science **209:** 457-463.

FRANCINO, M. P., L. CHAO, M. R. RILEY and H. OCHMAN, 1996 Asymmetries generated by transcription-coupled repair in enterobacterial genes Science **272:** 107 - 109.

FRANCINO, M. P., and H. OCHMAN, 1997 Strand asymmetries in DNA evolution. Trends in Genetics **13:** 240-245

FRANCINO, M. P., and H. OCHMAN, 2001 Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol. Biol. Evol. **18**.

FRIEDMAN, R., J. W. DRAKE and A. L. HUGHES, 2004 Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. Genetics **167**.

GALTIER, N., and L. DURET, 2007 Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics **23:** 273-277.

GALTIER, N., and J. R. LOBRY, 1997 Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes.

GARCIA-VALLVE, S., A. ROMEU and J. PALAU, 2000 Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. **10:** 1719-1725.

GARLAND, T., P. H. HARVEY and R. I. ANTHONY, 1992 Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systemat. Biol. **41:** 18-32.

GOFF, S. A., L. A. CASSON and A. L. GOLDBERG, 1984 Heat shock regulatory gene *htpR* influences rates of protein degradation and expression of the *lon* gene in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **81:** 6647-6651.

GOMAN, M., G. LANGSLEY, J. HYDE, N. YANKOVSKY, J. ZOLG *et al.*, 1982 The establishment of genomic DNA libraries for the human malaria parasite *Plasmodium falciparum* and identification of individual clones by hybridization. Mol. Biochem. Parasitol. **5:** 391-400.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity Nucl. Acids. Res. **10:** 7055-7074.

GOUY, M., C. GAUTIER, M. ATTIMONELLI, C. LANAVE and G. DI PAOLA, 1985 ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical design and usage. Comp. Appl. Biosci. **1:** 167–172.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE and R. MERCIER, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl. Acids. Res.

GREENACRE, M. J., 2007 Correspondence analysis in practice. Academic Press Inc . 115.

GROCOCK, R., and P. M. SHARP, 2001 Synonymous codon usage in *Cryptosporidium parvum*: identification of two distinct trends among genes. Int. J. Parisitol. **31:** 402-412.

GROCOCK, R. J., and P. M. SHARP, 2002 Synonymous codon usage in *Pseudomonas aeruginosa* PAO1. Gene **289:** 131-139.

GU, W., T. ZHOU and C. O. WILKE, 2010 A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comp. Biol. **6:** e1000664.

GUINDON, S., and O. GASCUEL, 2003 PhyML: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biol. **52:** 696-704.

HADDRILL, P. R., and B. C. CHARLESWORTH, 2008 Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. 4 **4:** 438 - 441

HALL, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series **41:** 95-98.

HARRISON, P. W., R. P. J. LOWER, N. K. D. KIM and J. P. W. YOUNG, 2010 Introducing the bacterial chromid: not a chromosome, not a plasmid. Trends in Microbiol **18:** 141-148.

HARRISON, R. J., and B. C. CHARLESWORTH, 2010 Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu-stricto* group of yeasts. Mol. Biol. Evol.**:** doi: 10.1093.

HARTL, D. L., E. N. MORIYAMA and S. A. SAWYER, 1994 Selection intensity for codon bias. Genetics **134:** 227-234.

HATFIELD, D. L., B. J. LEE and R. M. PIRTLE, 1992 Codon pair utilization bias in bacteria, yeast and mammals [in transfer RNA in protein synthesis]. CRC Press.

HENDRICKSON, E. L., R. KAUL, Y. ZHOU, D. BOVEE, P. CHAPMAN *et al.*, 2004 Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. J Bacteriol. **186:** 6956-6969.

HENRY, I., 2007 Evolution of codon usage in Bacteria. Ph.D. Thesis, University of Nottingham.

HENRY, I., and P. M. SHARP, 2006 Predicting gene expression level from codon usage bias. Mol Biol Evol **24:** 10-12.

HERBECK, J. T., D. P. WALL and J. J. WERNEGREEN, 2003 Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endomsymbiont *Wigglesworthia*. Microbiology **149:** 2585-2596.

HERSHBERG, R., and D. A. PETROV, 2009 General rules for optimal codon choice. PLoS Genetics **5:** e1000556.

HERSHBERG, R., and D. A. PETROV, 2010 Evidence that mutation is universally biased towards AT in bacteria. PloS Genet. **6:** e1001115.

HILDEBRAND, F., A. MEYER and A. EYRE-WALKER, 2010 Evidence of selection upon genomic GC-content in bacteria. PloS Genet. **6:** e1001107.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage in limits to artificial selection. Genet. Res. **8:** 269-294.

HOCHSTRASSER, M., 1995 Ubiquitin, proteasomes, and the regulation of intracellular protein degradation. Curr. Oppin. Cell Biol. **7:** 215-223

HOLLEY, R. W., J. APGAR, G. A. EVERETT, J. T. MADISON, M. MARQUISEE *et al.*, 1965 Structure of a ribonucleic acid. Science **147:** 1462-1465.

HORN, D., 2008 Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. BMC Genomics **9:** 2.

HUBER, H., M. J. HOHN, R. RACHEL, T. FUCHS, V. C. WIMMER *et al.*, 2002 A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature **417:** 63-67.

HURST, L. D., and A. R. MERCHANT, 2001 High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. R. Soc. Lond. **268** 493-497.

IKEMURA, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:** 389-409.

IKEMURA, T., 1982 Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. J. Mol. Biol. **158:** 573–597.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13-34.

IRWIN, B., J. D. HECK and H. G.W., 1995 Codon pair utilization biases influence translational elongation step times. J. Biol. Chem. **270:** 22801-22806.

JONES, W. J., M. J. B. PAYNTER and R. GUPTA, 1983 Characterization of *Methanococcus maripaludis* sp. nov., a new methanogen isolated from salt marsh sediment. Arch. Microbiol. **135:** 91-97.

KAISER, V. B., and B. C. CHARLESWORTH, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends in Genetics **25:** 9-12.

KANAYA, S., Y. YAMADA, M. KINOUCHI, Y. KUDO and T. IKEMURA, 2001 Codon usage and tRNA genes in Eukaryotes: correlation of codon usage dversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. **53:** 90–298.

KANAYA, S., Y. YAMADA, Y. KUDO and T. IKEMURA, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238:** 143-155.

KARLIN, S., J. MRAZEK, J. MA and L. BROCCHIERI, 2005 Predicted highly expressed genes in archaeal genomes. Proc. Natl. Acad. Sci. USA **102:** 7303-7308.

KERR, R. W., J. F. PEDEN and P. M. SHARP, 1997 Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Mol. Microbiol. **25:** 1177-1184.

KLAPPENBACH, J. A., J. M. DUNBAR and T. M. SCHMIDT, 2000 rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. **66:** 1328–1333.

KLASSON, L., and S. G. E. ANDERSSON, 2006 Strong asymmetric bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. Mol. Biol. Evol. **23:** 1031-1039.

KLOSTER, M., and C. TANG, 2008 SCUMBLE: a method for systematic and accurate detection of codon usage bias by maximum likelihood estimation. Nucleic Acids Res. **36:** 3819–3827.

KOSKI, L. B., R. A. MORTON and B. G. GOLDING, 2001 Codon bias and base composition are poor indicators of horizontally transferred genes. Mol. Biol. Evol. **18:** 404-412.

KUDLA, G., A. W. MURRAY, D. TOLLERVEY and J. B. PLOTKIN, 2009 Coding-sequence determinants of gene expression in *Escherichia coli*. Science **324:** 255-258.

LAFAY, B., J. C. ATHERTON and P. SHARP, 2000 Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology **146:** 851-860.

LAFAY, B., A. T. LLOYD, M. J. MCLEAN, K. M. DEVINE and P. M. SHARP, 1999 Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucl. Acids. Res. **27:** 1642-1649.

LAFAY, B., and P. M. SHARP, 1999 Synonymous codon usage variation among *Giardia lamblia* genes and isolates. Mol. Biol. Evol. **16:** 1484-1495.

LAWRENCE, J. G., and H. OCHMAN, 1997 Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:** 383-397.

LECOMPTE, O., R. RIPP, J. C. THIERRY, D. MORAS and O. POCH, 2002 Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucl. Acids. Res. **30:** 5382-5390.

LIN, S., and I. ZABIN, 1972 Beta-galactosidase: rates of synthesis and degradation of incomplete chains. J. Biol. Chem. **247:** 2205.

LIND, P. A., and D. I. ANDERSSON, 2008 Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. USA **105:** 17878-17883.

LINDAHL, T., and B. NYBERG, 1974 Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry **13** 3405-3410.

LINDÅS, A. C., E. A. KARLSSON, M. T. LINDGREN, T. J. G. ETTEMA and R. BERNANDER, 2008 A unique cell division machinery in the archaea. Proc. Natl. Acad. Sci. USA **105:** 18942-18946

LLOYD, A. T., and P. M. SHARP, 1991 Codon usage in *Aspergillus nidulans*. Mol. & Gen. Genet.

LLOYD, A. T., and P. M. SHARP, 1993 Synonymous codon usage in *Kluyveromyces lactis*. Yeast **9:** 1219 - 1228.

LOBRY, J. R., 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:** 660-665.

LOBRY, J. R., and D. CHESSEL, 2003 Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J. Appl. Genet. **44:** 235-261.

LOBRY, J. R., and A. NECSULEA, 2006 Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. Gene **385:** 128-136.

LOBRY, J. R., and N. SUEOKA, 2002 Asymmetric directional mutation pressures in Bacteria. Genome Biology **3**.

LOVMAR, M., and M. EHRENBERG, 2006 Rate, accuracy and cost of ribosomes in bacterial cells Biochimie **88:** 951-961.

LOWE, T. M., and S. R. EDDY, 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids. Res. **25:** 955-964.

LUNDGREN, M., A. ANDERSSON, L. CHEN, P. NILSSON and R. BERNANDER, 2004 Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination. Proc. Natl. Acad. Sci. USA **101:** 7046-7051.

LYNCH, M., 2007 The origins of genome architecture. MA, USA: Sinauer Associates**.**

LYNCH, M., B. KOSKELLA and S. SCHAACK, 2006 Mutation pressure and the evolution of organelle genomic architecture Science **311:** 1727 - 1730.

LYNN, D. J., G. A. C. SINGER and D. A. HICKEY, 2002 Synonymous codon usage is subject to selection in thermophilic bacteria. Nucl. Acids. Res. **30:** 4272-4277.

MAKAROVA, K. S., L. ARAVIND, N. V. GRISHIN, I. B. ROGOZIN and E. V. KOONIN, 2002 A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis Nucl. Acids. Res. **30:** 482-496.

MANCERA, E., R. BOURGON, A. BROZZI, W. HUBER and L. STEINMETZ, 2008 High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature **454:** 479-485.

MANN, S., and Y.-P. P. CHEN, 2010 Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics **95:** 7–15.

MARAIS, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends in Genetics **19**: 330-338.

MARTÍN-GALIANO, A. J., J. M. WELLS and A. G. DE LA CAMPA, 2004 Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. Microbiology **150:** 2313-2325.

MARTINDALE, D. W., 1989 Codon usage in *Tetrahymena* and other ciliates. J. Protozool. **36:** 29-34.

MASIDE, X., A. W. LEE and B. CHARLESWORTH, 2004 Selection on codon usage in *Drosophila americana*. Curr. Biol. **14:** 150-154.

MAYNARD SMITH, J., and N. H. SMITH, 1996 Site-specific codon bias in Bacteria. Genetics **142:** 1037-1043.

MCINERNEY, J. O., 1997 Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microbial & Comparative Genomics **2:** 89-97.

MCLEAN, M., K. H. WOLFE and K. M. DEVINE, 1998 Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. **47:** 691–696.

MCVEAN, G. A. T., and B. CHARLESWORTH, 1999 A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res., Camb. **74:** 145-158.

MEDIGUE, C., P. ROUXEL, P. VIGIER, A. HENAUT and A. DANCHIN, 1991 Evidence for horizontal gene transfer in *Escherichia coli*. J. Mol. Biol. **222:** 851-856.

MIKKOLA, R., and C. G. KURLAND, 1992 Selection of laboratory wild-type phenotype from natural isolates of *Escherichia coli* in chemostats. Mol. Biol. Evol. **9:** 394-402.

MIRA, A., H. OCHMAN and N. A. MORAN, 2001 Deletional bias and the evolution of bacterial genomes. Trends. Genet. **17:** 589-596.

MORAN, N. A., H. J. MCLAUGHLIN and R. SOREK, 2009 The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Nature **323:** 379-382.

MORIYAMA, E. N., and J. R. POWELL, 1997 Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45:** 514–523.

MUSTO, H., H. NAYA, A. ZAVALA, H. ROMERO, F. ALVAREZ-VALIN *et al.*, 2004 Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS **573:** 73-77.

MUSTO, H., H. NAYA, A. ZAVALA, H. ROMERO, F. ALVAREZ-VALIN *et al.*, 2006 Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochemical and Biophysical Research Communications **347:** 1-3.

MUSTO, H., H. ROMERO and A. ZAVALA, 2003 Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*. Microbiology **149** 855.

MUTO, A., and S. OSAWA, 1987 The guanine and cytosine content of the genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA **84:** 166–169.

MUTO, A., F. YAMAO and Y. KAWAUCHI, 1985 Codon usage in *Mycoplasma capricolum*. Proc. Jap. Acac. B. **61:** 12-15.

NAGYLAKI, T., 1983 Evolution of a finite population under gene conversion. Proc. Nat. Acad. Sci. USA **80:** 6278-6681.

NAYA, H., H. ROMERO, N. CARELS, A. ZAVALA and H. MUSTO, 2001 Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii* FEBs Lett. **501:** 127-130.

NAYA, H., H. ROMERO, A. ZAVALA, B. ALVAREZ and H. MUSTO, 2002 Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J. Mol. Evol. **55:** 260-264.

NECSULEA, A., and J. R. LOBRY, 2007 A new method for assessing the effect of replication on DNA base composition asymmetry. Mol. Biol. Evol.

NEI, M., and D. GRAUR, 1984 Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17:** 73-118.

NOVEMBRE, J. A., 2002 Accounting for background nucleotide composition when measuring codon usage bias. Mol. Biol. Evol. **19:** 1390-1394.

OCHMAN, H., J. G. LAWRENCE and E. A. GROISMAN, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature **405:** 299-304.

OCHMAN, H., and A. C. WILSON, 1987 Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes J. Mol. Evol. **26:** 74-86.

OLLER, A. R., I. J. FIJALKOWSKA, R. L. DUNN and R. M. SCHAAPER, 1992 Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli*. Proc. Nat. Acad. Sci. USA **89:** 11036-11040.

ORTEGO, B. C., J. J. WHITTENTON, H. L. SHIAO-CHUN TU and R. C. WILLSON, 2007 In vivo translational inaccuracy in *Escherichia coli*: missense reporting using extremely low activity mutants of *Vibrio harveyi* luciferase. Biochemistry **46**.

PARADIS, E., J. CLAUDE and K. STRIMMER, 2004 APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20:** 289-290.

PARKER, J., 1989 Errors and alternatives in reading the universal genetic code. Microbiol. Rev. **53:** 273-298.

PARKER, J., and J. PRECUP, 1986 Mistranslation during phenylalanine starvation. Mol. Gen. Genetic. **204:** 70-74.

PEDERSEN, S., P. L. BLOCH, S. REEH and F. C. NEIDHARDT, 1978 Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. Cell **14:** 179-190.

PEIXOTO, L., V. FERNANDEZ and H. MUSTO, 2004 The effect of expression levels on codon usage in *Plasmodium falciparum*. Parasitology **128:** 245-251.

PERRIERE, G., and J. THIOULOUSE, 2002 Use and misuse of correspondence analysis in codon usage studies. Nucl. Acids. Res. **30:** 4548-4555.

PHIPPS, B. M., D. TYPKE, R. HEGERL, S. VOLKER, A. HOFFMAN *et al.*, 1993 Structure of a molecular chaperone from a thermophilic archaebacterium. Nature **361:** 475-447.

PINE, M. J., 1965 Heterogeneity of protein turnover in *Escherichia coli*. Biochimica *et* Biophysica Acta **104:** 439-456.

PRECUP, J., and J. PARKER, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:** 11351–11355.

PUIGBO, P., A. PASAMONTES and S. GARCIA-VALLVE, 2008 Gaining and losing the thermophilic adaptation in prokaryotes. Trends in Genetics **24:** 10-14.

ROCHA, E. P. C., 2004 Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation. Genome Research **14:** 2279-2286.

ROCHA, E. P. C., M. TOUCHON and E. J. FEIL, 2006 Similar compositional biases are caused by very different mutational effects. Genome Res. **16:** 1537-1547.

ROMANO, V., A. NAPOLI, V. SALERNO, A. VALENTI, M. ROSSI *et al.*, 2007 Lack of strand-specific repair of UV-induced DNA lesions in three genes of the archaeon *Sulfolobus solfataricus*. J. Mol. Biol. **365:** 921-929.

ROMERO, H., A. ZAVALA and H. MUSTO, 2000 Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*. Gene **242:** 307-311.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406-425.

SALIM, H. M. W., K. L. RING and A. R. O. CAVALCANTI, 2008 Patterns of codon usage in two ciliates that reassign the genetic code: *Tetrahymena thermophila* and *Paramecium tetraurelia*. Protist. **159:** 283-298.

SÉMON, M., D. MOUCHIROUD and L. DURET, 2005 Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Human. Mol. Genetic. **14:** 421-427.

SHARP, P. M., 1991 Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: Codon usage, map position, and concerted evolution. J. Mol. Evol. **33:** 23-33.

SHARP, P. M., E. BAILES, R. J. GROCOCK, J. F. PEDEN and R. E. SOCKETT, 2005 Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. **33:** 1141-1153.

SHARP, P. M., E. COWE, D. HIGGINS, D. SHIELDS, K. WOLFE *et al.*, 1988 Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res **16:** 8207-8211.

SHARP, P. M., and K. M. DEVINE, 1989 Codon usage and gene-expression level in *Dictyostelium discoideum* - highly expressed genes do prefer optimal codons. Nucleic Acids Res. **17:** 5029-5039.

SHARP, P. M., L. R. EMERY and K. ZENG, 2010 Forces that influence the evolution of codon bias. Phil. Trans. R. Soc. **365:** 1203-1212.

SHARP, P. M., and W.-H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28-38.

SHARP, P. M., and W. H. LI, 1987a The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucl. Acids. Res. **15:** 1281-1295.

SHARP, P. M., and W. H. LI, 1987b The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222-230.

SHARP, P. M., and A. T. LLOYD, 1993 Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res. **21:** 179–183.

SHARP, P. M., T. M. F. TUOHY and K. R. MOSURSKI, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes Nucleic Acids Res. **14:** 5125-5143.

SHIELDS, D. C., 1990 Switches in species-specific codon preferences: The influence of mutation biases. J. Mol. Evol. **31:** 71-80.

SHIELDS, D. C., and P. M. SHARP, 1987 Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. **15:** 8023-8040.

SHIN-ICHI, A., and T. IKEMURA, 1986 Diversity in G+C content at the third position of codons in vertebrate genes and its cause Nucleic Acids Res. **14:** 6345-6355.

SØRENSEN, M. A., C. G. KURLAND and S. PEDERSEN, 1989 Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. **207:** 365-377.

SØRENSEN, M. A., and S. PEDERSEN, 1991 Absolute in vivo translation rates of individual codons in *Escherichia coli*: the two glutamic acid codons GAA and GAG are translated with a threefold difference in rate J. Mol. Biol. **222:** 265-280.

ST JOHN, A. C., and A. L. GOLDBERG, 1978 Effects of reduced energy production on protein degradation, guanosine tetraphosphate, and RNA synthesis in *Escherichia coli*. J. Biol. Chem. **253:** 2705 - 2711

STENICO, M., T. ANDREW and P. M. SHARP, 1994 Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. **22:** 437–2446.

STEVENSON, S. B., and T. M. SCHMIDT, 2004 Life history implications of rRNA gene copy number in *Escherichia coli*. Appl. Environ. Microbiol. **70:** 6670-6677.

STOLETZKI, N., and A. EYRE-WALKER, 2007 Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. Mol. Biol. Evol. **24:** 374-381.

SUBRAMANIAN, S., 2008 Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics **178:** 2429–2432.

SUOEKA, N., 1962 On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48:** 582-592.

SUZUKI, H., C. J. BROWN, L. J. FORNEY and E. M. TOP, 2008 Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res. **15:** 357-365.

TAKEMOTO, K., and T. AKUTSU, 2008 Origin of structural difference in metabolic networks with respect to temperature. BMC Systems Biology **2:** 82.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids. Res. **22:** 4673-4680.

TOFT, C., and M. A. FARES, 2009 Selection for translational robustness in *Buchnera aphidicola*, endosymbiotic bacteria of aphids. Mol. Biol. Evol. **26:** 743-751.

TULLER, T., A. CARMI, K. VESTSIGIAN, S. NAVON, S. DORFAN *et al.*, 2010 An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell **141:** 344-354.

TULLER, T., Y. Y. WALDMAN, M. KUPIEC and E. RUPPINC, 2009 Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. USA**:** (in press).

VAN BOGELEN, R. A., P. SANKAR, R. L. CLARK, J. A. BOGAN and F. NEIDHARDT, 1992 Gene-protein database of *Escherichia coli* K-12: edition 5. Electrophoresis **13:** 1014-1054.

VAN PASSEL, M. W. J., A. BART, A. C. M. LUYF, A. H. C. VAN KAMPEN and A. VAN DER ENDE, 2006 Compositional discordance between prokaryotic plasmids and host chromosomes. BMC Genomics **7:** 26.

VARANI, G., and W. H. MCCLAIN, 2000 The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. EMBO reports **1:** 18.

VIEIRA-SILVA, S., and E. P. C. ROCHA, 2010 The systemic imprint of growth and Its uses in ecological (meta)genomics. PLoS Genetics **6:** e1000808.

WARNECKE, T., and L. D. HURST, 2010 GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution Mol. Sys. Biol. **6:** 340.

WEN, J. D., L. LANCASTER, C. HODGES, A. C. ZERI, S. H. YOSHIMURA *et al.*, 2008 Following translation by single ribosomes one codon at a time. Nature **452:** 598-603.

WERNEGREEN, J. J., and D. J. FUNK, 2004 A neutral explanation for extreme base composition of an endosymbiont genome. J. Mol. Evol. **59:** 849-858.

WHELAN, S., and N. GOLDMAN, 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18:** 691-699.

WILKE, C. O., and D. A. DRUMMOND, 2006 The population genetics of translational robustness. Genetics **173:** 473-481.

WORNING, P., L. J. JENSEN, P. F. HALLIN, H. H. STAERFELDT and D. W. USSERY, 2006 Origin of replication in circular prokaryotic chromosomes. Environ. Microbiol. **8:** 353-361.

WRIGHT, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23-29.

WRIGHT, S., 1932 The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc.6th Int. Congress Genetics **1:** 356-366.

XIA, Q., E. L. HENDRICKSON, Y. ZHANG, T. WANG, F. TAUB *et al.*, 2006 Quantitative proteomics of the archaeon *Methanococcus maripaludis* validated by microarray analysis and real time PCR Mol. & Cell. Proteom. **5:** 868-881.

XIA, X., 1995 Body temperature, rate of biosynthesis, and evolution of genome size. Mol. Biol. Evol. **12:** 834-842.

ZAVALA, A., H. NAYA, V. SABBIA, R. PIOVANI and H. MUSTO, 2005 Genomic GC content prediction in prokaryotes from a sample of genes. Gene **357:** 137-143.

ZHOU, T., M. WEEMS and C. O. WILKE, 2009 Translationally optimal codons associate with structurally sensitive sites in proteins. Mol. Biol. Evol. **26:** 1571–1580.

APPENDICES

Appendix A Species of Archaea analysed

[a]Code used to identify each species throughout this report
[b]Published genome sequence reference
[c]Year genome sequence was published
[d]GenBank accession number

Appendix B Generation time and optimal growth temperature sources

[a]Minimal generation time for the 67 species of Archaea analysed
[b]Reference for minimal generation time
[c]Optimal growth temperature
[d]Reference for minimal generation time if different from [b] or unavailable from http://www.dsmz.de

Appendix C Oxygen requirement and habitat status of Archaea

According to NCBI classification for 67 species. The habitat classification is somewhat limited since a species could be both host-associated and specialised (for example) and yet only one habitat is assigned per species.

Appendix D Species coordinates upon within-group correspondence analysis axes (WCA)

In analyses of (i) genome-wide codon usage (CU) and (ii) codon usage across 20 highly expressed genes among 67 species of Archaea. The first four axes in each analysis are (HARRISON *et al.* 2010)shown.

Appendix E Optimal codons among Archaea

[a] The strength of selected codon usage bias as in Chapter 4
[b]Average genome-wide G+C content across synonymously variable third codon positions
[c] Optimal codon identity; colours indicate third optimal codon position for those significant at $p < 0.01$ level of significance; grey indicates those at $p > 0.05$

Appendix F Numbers of tRNA anticodons decoding synonymous codons in Archaea

[a]Numbers obtained from the tRNA-Scan database (LOWE and EDDY 1997) for 67 species analysed.

Appendix G Estimates of the strength of selection (*S*) adapted for each amino acid group

Across 67 species of Archaea analysed. Grey indicates values exclude from analyses either because an optimal codon for that amino acid, or the presence of translational selection was not detected.

Species	Code[a]	Reference[b]	Year[c]	Accession[d]
Euryarchaeaota				
Archaeoglobales				
Archaeoglobus fulgidus	Arcful	Nature. 390:364	1997	AE000782
Archaeoglobus profundus	Arcpro	Unpublished		CP001857
Ferroglobus placidus	Ferpla	Unpublished		CP001899
Halobacteriales				
Haloarcula marismortui	Halmar	Genome Res. 14:2221	2004	AY596297
Halobacterium salinarum	Halsal	Genomics. 91(4):335	2008	AM774415
Halorubrum lacusprofundi	Hallac	Unpublished		CP001365
Halomicrobium mukohataei	Halmuk	Unpublished		CP001688
Halorhabdus utahensis	Haluta	Unpublished		CP001687
Haloterrigena turkmenica	Haltur	Unpublished		CP001860
Haloquadratum walsbyi	Halwal	BMC Genomics. 7:169	2006	AM180088
Natrialba magadii	Natmag	Unpublished		CP001932
Natronomonas pharaonis	Natpha	Genome Res.15:1336	2005	CR936257
Methanobacteriales				
Methanobacterium thermoautotrophicus	Metthe	J. Bacteriol. 179:7135	1997	AE000666
Methanobrevibacter ruminantium	Metrum	PLoS One. 5(1):e8926	2010	CP001719
Methanobrevibacter smithii	Metsmi	PNAS. 104:10643	2007	CP000678
Methanosphaera stadtmanae	Metsta	J. Bacteriol. 188:642	2006	CP000102
Methanopyrales				
Methanopyrus kandleri	Metkan	PNAS. 99:4644	2002	AE009439
Methanococcales				
Methanocaldococcus jannaschii	Metjan	Science. 273:1058	1996	L77117
Methanocaldococcus fervens	Metfer	Unpublished		CP001696
Methanocaldococcus vulcanius	Metvul	Unpublished		CP001787
Methanococcus aeolicus	Metaeo	Unpublished		CP000743
Methanococcus maripaludis S2	MetmarS2	J. Bacteriol. 186:6956	2004	BX950229
Methanococcus vannielii	Metvan	Unpublished		CP000742
Methanosarcinales				
Methanococcoides burtonii	Metbur	Unpublished		CP000300
Methanosaeta thermophila PT	MetthePT	Unpublished		CP000477
Methanosarcina acetivorans	Metace	Genome Res. 12:532	2002	AE010299
Methanosarcina barkeri fusaro	Metbak	J. Bacteriol. 188:7922	2006	CP000099
Methanosarcina mazei	Metmaz	JMMB. 4:453	2002	AE008384
Rice Cluster I MRE50	RicClu	Science. 313:370	2006	AM114193
Methanomicrobiales				
Methanocorpusculum labreanum	Metlab	PLoS One. 4(6):e5797	2009	CP000559
Methanoculleus marisnigri JR1	MetmarJR	PLoS One. 4(6):e5797	2009	CP000562
Methanoregula boonei	Metboo	Unpublished		CP000780
Methanospirillum hungatei	Methun	Unpublished		CP000254
Methanocella paludicola	Metpa2	Unpublished		AP011532
Methanosphaerula palustris	Metpal	Unpublished		CP001338

A

Species	Code[a]	Reference[b]	Year[c]	Accession[d]
Euryarchaeaota				
Thermococcales				
Pyrococcus abyssi	Pyraby	Mol. Microbiol. 47:1495	2003	AL096836
Pyrococcus furiosus	Pyrfur	Meth. Enz. 330:134	2001	AE009950
Pyrococcus horikoshii	Pyrhor	DNA Res. 5:55	1998	BA000001
Thermococcus gammatolerans	Thegam	Genome Biol. 10(6):R70	2009	CP001398
Thermococcus kodakarensis	Thekod	Genome Res. 15:352	2005	AP006878
Thermococcus onnurineus	Theonu	J. Bacteriol. 190(22):7491	2008	CP000855
Thermococcus sibiricus	Thesib	Appl. Env. Microbiol. 75(13):4580	2009	CP001463
Thermoplasmatales				
Picrophilus torridus	Pictor	PNAS. 101:9091	2004	AE017261
Thermoplasma acidophilum	Theaci	Nature. 407:508	2000	AL139299
Thermoplasma volcanium	Thevol	PNAS. 97:14257	2000	BA000011
Aciduliprofundum boonei	Aciboo	Unpublished		CP001941
Crenarchaeota				
Desulfurococcales				
Aeropyrum pernix	Aerper	DNA Res. 6:83	1999	BA000002
Hyperthermus butylicus	Hypbut	Archaea. 2:127	2007	CP000493
Ignicoccus hospitalis	Ignhos	Genome Biol. 9(11):R158	2008	CP000816
Staphylothermus marinus	Stamar	BMC Genomics. 10:145	2009	CP000575
Desulfurococcus kamchatkensis	Deskam	J. Bacteriol. 191(7):2371	2009	CP001140
Thermoproteales				
Caldivirga maquilingensis	Calmaq	Unpublished		CP000852
Pyrobaculum aerophilum	Pyraer	PNAS. 99:984	2002	AE009441
Pyrobaculum arsenaticum	Pyrars	Unpublished		CP000660
Pyrobaculum calidifontis	Pyrcal	Unpublished		CP000561
Pyrobaculum islandicum	Pyrisl	Unpublished		CP000504
Thermoproteus neutrophilus	Theneu	Unpublished		CP001014
Thermofilum pendens	Thepen	J. Bacteriol. 190(8):2957	2008	CP000505
Sulfolobales				
Metallosphaera sedula	Metsed	Appl. Env. Microbiol. 74(3):682	2008	CP000682
Sulfolobus acidocaldarius	Sulaci	J. Bacteriol.187:4992	2005	CP000077
Sulfolobus islandicus	Sulisl	PNAS. 106(21):8605		CP001399
Sulfolobus solfataricus	Sulsol	PNAS. 98:7835	2001	AE006641
Sulfolobus tokodaii	Sultok	DNA Res. 8:123	2001	BA000023
Thaumarchaeota				
Nitrosopumilus maritimus	Nitmar	PNAS. 107(19):8818	2010	CP000866
Cenarchaeum symbiosum A	Censym	PNAS. 103:18296	2006	DP000238
Korarchaeota				
Korarchaeum cryptofilum	Korcry	PNAS. 105(23):8102	2008	CP000968
Nanoarchaeota				
Nanoarchaeum equitans	Natequ	PNAS. 100:12984	2003	AE017199

A

Code[a]	GenT[a]	Reference[b]		Temp[c]	Reference[d]
Euryarchaeaota					
Archaeoglobales					
Arcful	4.0	Nature. 390:364	1997	83	
Arcpro	4.0	Syst. Appl. Microbiol. 13:24	1990	82	
Ferpla	2.8	Arch. Microbiol. 167:19	1997	85	
Halobacteriales					
Halmar	10.0	FEMS. 241:21	2004	37	
Halsal	4.0	BMC Genomics. 8:415	2007	37	
Hallac	11.1	Sys. Appl. Microbiol. 11:20	1988	31	
Halmuk	- -	- -	- -	45	
Haluta	- -	- -	- -	50	
Haltur	1.5	J. Bacteriol. 187:923	2005	51	
Halwal	24.0	Env. Microbiol. 6:1287	2004	37	
Natmag	12.0	Lett. Appl. Microbiol. 44(6):637	2007	37	
Natpha	2.1	J. Bacteriol. 187(3):923	2005	44	
Methanobacteriales					
Metthe	1.8	Arch. Microbiol. 127: 59	1980	65	
Metrum	16.8	IJSEM. 57(3): 450	2007	37	
Metsmi	4.4	Curr. Micobiol. 40:176	2000	37	
Metsta	5.3	FEMS. 60:266	2007	37	
Methanopyrales					
Metkan	0.8	Env. Microbiol. 7:47	2005	98	
Methanococcales					
Metjan	0.5	Extremophiles. 11:495	2007	83	
Metfer	- -	- -	- -	- -	
Metvul	0.8	IJSEM. 49:583	1999	80	
Metaeo	2.0	IJSEM. 56:1525	2006	46	
MetmarS2	2.3	Arch. Microbiol. 135:91	1983	37	
Metvan	5.8	Arch. Microbiol. 141(2):133	1986	38	
Methanosarcinales					
Metbur	20.0	Mol. Microbiol. 53:309	2004	22	
MetthePT	- -	- -	- -	65	
Metace	24.0	PNAS. 101:16929	2004	38	
Metbak	6.9	FEMS. 60:266	2007	34	
Metmaz	10.0	Env. Microbiol. 7:47	2005	34	
RicClu	- -	Appl. Env. Microbiol. 73:4326	2007	37	
Methanomicrobiales					
Metlab	13.0	IJSEM. 39(1):10	1989	37	
Metmar	- -	- -	- -	40	IJSEM 40:117 (1990)
Metboo	48.0	Nature. 442:4810	2005	37	
Methun	14.0	Arch. Microbiol. 141(2):133	1986	37	
Metpa2	101.0	IJSEM. 58:929	2008	37	
Metpal	28.0	Appl. Env. Microbiol. 74(7):2059	2008	30	

Code[a]	GenT[a]	Reference[b]	Year	Temp[c]	Reference[d]
Euryarchaeaota					
Thermococcales					
Pyraby	0.6	Extremophiles 2:123	1998	96	
Pyrfur	0.6	Extremophiles 2:123	1998	99	
Pyrhor	0.5	Extremophiles 2:123	1998	95	
Thegam	1.6	IJSEM. 53:847	2003	88	
Thekod	0.7	J. Biotech. 116:271	2005	85	
Theonu	- -	- -	- -	80	
Thesib	0.7	Extremophiles. 5:85	2001	78	
Thermoplasmatales					
Pictor	6.0	J. Bactriol. 177:7050	1995	60	
Theaci	2.5	Env.Microbiol. 7:47	2005	58	
Thevol	2.5	Env.Microbiol. 7:47	2005	60	
Aciboo	- -	- -	- -	- -	
Crenarchaeota					
Desulfurococcales					
Aerper	3.3	Env.Microbiol. 7:47	2005	90	
Hypbut	2.0	Env.Microbiol. 7:47	2005	99	
Ignhos	1.0	J. Bacteriol. 190(5): 1743	2008	90	
Stamar	- -	- -	- -	88	
Deskam	- -	- -	- -	82	
Thermoproteales					
Calmaq	8.0	IJSEM. 49(3):1157	1999	83	
Pyraer	3.0	Env.Microbiol. 7:47	2005	98	
Pyrars	3.7	Appl. Env. Microbiol. 67(12):5568	2001	95	
Pyrcal	5.5	Archaea. 1:113	2002	93	
Pyrisl	2.7	J. Bactriol. 188:4350	2006	98	
Theneu	- -	- -	- -	85	
Thepen	- -	- -	- -	88	
Sulfolobales					
Metsed	4.5	Extremophiles. 5(4):241	2001	65	
Sulaci	5.7	Env.Microbiol. 7:47	2005	70	
Sulisl	- -	- -	- -	75	
Sulsol	6.0	Env.Microbiol. 7:47	2005	78	
Sultok	6.0	Env.Microbiol. 7:47	2005	75	
Thaumarchaeota					
Nitmar	30.0	Nature. 437:543	2005	28	
Censym	- -	- -	- -	10	PNAS. 93:6241 (1996)
Korarchaeota					
Korcry	- -	- -	- -	83	Extremeophil. 4:61 (2000)
Nanoarchaeota					
Nanequ	0.75	J. Bacteriol. 190(5):1743	2008	90	

Species	Oxygen requirement	Habitat
Euryarchaeaota		
Archaeoglobales		
Archaeoglobus fulgidus	Anaerobic	Aquatic
Archaeoglobus profundus	Anaerobic	Specialized
Ferroglobus placidus	Anaerobic	Specialized
Halobacteriales		
Haloarcula marismortui	Aerobic	Specialized
Halobacterium salinarum	Anaerobic	Specialized
Halorubrum lacusprofundi	Aerobic	Aquatic
Halomicrobium mukohataei	Facultative	Specialized
Halorhabdus utahensis	Aerobic	Terrestrial
Haloterrigena turkmenica	Aerobic	Specialized
Haloquadratum walsbyi	Aerobic	Aquatic
Natrialba magadii	Aerobic	Specialized
Natronomonas pharaonis	Microaerophilic	Aquatic
Methanobacteriales		
Methanobacterium thermoautotrophicus	Anaerobic	Specialized
Methanobrevibacter ruminantium	Anaerobic	- -
Methanobrevibacter smithii	Anaerobic	Multiple
Methanosphaera stadtmanae	Anaerobic	Host-associated
Methanopyrales		
Methanopyrus kandleri	Anaerobic	Specialized
Methanococcales		
Methanocaldococcus jannaschii	Anaerobic	Aquatic
Methanocaldococcus fervens	Anaerobic	Specialized
Methanocaldococcus vulcanius	Anaerobic	Specialized
Methanococcus aeolicus	Anaerobic	Aquatic
Methanococcus maripaludis S2	Anaerobic	Aquatic
Methanococcus vannielii	Anaerobic	Aquatic
Methanosarcinales		
Methanococcoides burtonii	Anaerobic	Aquatic
Methanosaeta thermophila PT	Anaerobic	Aquatic
Methanosarcina acetivorans	Anaerobic	Aquatic
Methanosarcina barkeri fusaro	Anaerobic	Multiple
Methanosarcina mazei	Anaerobic	Multiple
Rice Cluster I MRE50	- -	Host-associated
Methanomicrobiales		
Methanocorpusculum labreanum	Anaerobic	Aquatic
Methanoculleus marisnigri JR1	Anaerobic	Aquatic
Methanoregula boonei	Anaerobic	Terrestrial
Methanospirillum hungatei	Anaerobic	Multiple
Methanocella paludicola	Anaerobic	- -
Methanosphaerula palustris	Anaerobic	Specialized

Species	Oxygen requirement	Habitat
Euryarchaeaota		
Thermococcales		
Pyrococcus abyssi	Anaerobic	Aquatic
Pyrococcus furiosus	Anaerobic	Aquatic
Pyrococcus horikoshii	Anaerobic	Aquatic
Thermococcus gammatolerans	Anaerobic	Specialized
Thermococcus kodakarensis	Anaerobic	Specialized
Thermococcus onnurineus	Anaerobic	Terrestrial
Thermococcus sibiricus	Anaerobic	- -
Thermoplasmatales		
Picrophilus torridus	Aerobic	Specialized
Thermoplasma acidophilum	Facultative	Specialized
Thermoplasma volcanium	Facultative	Specialized
Aciduliprofundum boonei	Anaerobic	Aquatic
Crenarchaeota		
Desulfurococcales		
Aeropyrum pernix	Aerobic	Specialized
Hyperthermus butylicus	Anaerobic	Aquatic
Ignicoccus hospitalis	Anaerobic	Aquatic
Staphylothermus marinus	Anaerobic	Specialized
Desulfurococcus kamchatkensis	Anaerobic	Aquatic
Thermoproteales		
Caldivirga maquilingensis	Microaerophilic	Specialized
Pyrobaculum aerophilum	Facultative	Aquatic
Pyrobaculum arsenaticum	Anaerobic	Aquatic
Pyrobaculum calidifontis	Facultative	Specialized
Pyrobaculum islandicum	Anaerobic	Specialized
Thermoproteus neutrophilus	Anaerobic	Specialized
Thermofilum pendens	Anaerobic	Specialized
Sulfolobales		
Metallosphaera sedula	Aerobic	Specialized
Sulfolobus acidocaldarius	Aerobic	Specialized
Sulfolobus islandicus	Aerobic	Specialized
Sulfolobus solfataricus	Aerobic	Specialized
Sulfolobus tokodaii	Aerobic	Specialized
Thaumarchaeota		
Nitrosopumilus maritimus	Aerobic	Aquatic
Cenarchaeum symbiosum A	Anaerobic	Specialized
Korarchaeota		
Korarchaeum cryptofilum	Anaerobic	Specialized
Nanoarchaeota		
Nanoarchaeum equitans	Anaerobic	Host-associated

Species	WCA of CU genome-wide				WCA of highly expressed CU			
	1	2	3	4	1	2	3	4
Euryarchaeaota								
Archaeoglobales								
Archaeoglobus fulgidus	0.01	0.21	-0.12	-0.06	0.10	-0.13	0.14	0.17
Archaeoglobus profundus	-0.27	0.16	0.00	-0.13	-0.18	-0.13	0.09	0.15
Ferroglobus placidus	-0.14	0.12	-0.02	-0.07	-0.06	-0.13	0.01	0.25
Halobacteriales								
Haloarcula marismortui	0.56	-0.18	0.06	-0.04	0.69	0.28	-0.13	0.02
Halobacterium salinarum	0.78	-0.13	0.18	0.09	0.81	0.20	-0.27	0.02
Halorubrum lacusprofundi	0.77	-0.16	0.17	-0.02	0.82	0.23	-0.24	0.04
Halomicrobium mukohataei	0.74	-0.18	0.12	-0.03	0.78	0.28	-0.16	0.01
Halorhabdus utahensis	0.60	-0.20	0.07	-0.05	0.74	0.26	-0.18	0.00
Haloterrigena turkmenica	0.78	-0.19	0.16	-0.06	0.78	0.28	-0.17	0.05
Haloquadratum walsbyi	-0.17	-0.27	-0.11	-0.23	-0.15	0.26	0.09	-0.34
Natrialba magadii	0.55	-0.19	0.08	-0.13	0.65	0.31	-0.07	0.03
Natronomonas pharaonis	0.58	-0.20	0.08	-0.05	0.70	0.30	-0.15	0.01
Methanobacteriales								
Methanobacterium thermoautotrophicus	0.00	0.20	-0.22	-0.06	0.08	-0.10	0.24	0.07
Methanobrevibacter ruminantium	-0.49	-0.10	-0.02	0.01	-0.57	0.32	-0.03	0.11
Methanobrevibacter smithii	-0.70	-0.27	0.07	0.04	-0.76	0.32	-0.06	0.02
Methanosphaera stadtmanae	-0.80	-0.25	0.11	0.01	-0.84	0.30	-0.05	-0.08
Methanopyrales								
Methanopyrus kandleri	0.46	0.00	0.01	-0.03	0.62	0.06	-0.06	-0.04
Methanococcales								
Methanocaldococcus jannaschii	-0.65	-0.10	0.18	-0.01	-0.66	0.13	-0.09	0.27
Methanocaldococcus fervens	-0.65	-0.10	0.17	0.00	-0.67	0.14	-0.10	0.29
Methanocaldococcus vulcanius	-0.60	-0.14	0.13	0.02	-0.65	0.11	-0.12	0.14
Methanococcus aeolicus	-0.66	-0.21	0.17	0.12	-0.73	0.16	-0.10	0.07
Methanococcus maripaludis S2	-0.58	-0.28	0.02	0.07	-0.62	0.32	-0.02	0.33
Methanococcus vannielii	-0.64	-0.27	0.07	0.08	-0.64	0.28	-0.04	0.19
Methanosarcinales								
Methanococcoides burtonii	-0.29	-0.09	-0.18	-0.03	-0.22	0.14	0.33	-0.11
Methanosaeta thermophila PT	0.19	0.20	-0.15	-0.02	0.32	-0.19	0.08	-0.06
Methanosarcina acetivorans	-0.15	-0.10	-0.18	0.09	0.03	0.17	0.21	-0.01
Methanosarcina barkeri fusaro	-0.29	-0.12	-0.15	0.06	-0.12	0.15	0.20	-0.10
Methanosarcina mazei	-0.21	-0.09	-0.20	0.09	-0.03	0.15	0.25	-0.02
Rice Cluster I MRE50	0.38	0.02	-0.10	0.12	0.52	0.08	0.12	0.09
Methanomicrobiales								
Methanocorpusculum labreanum	0.13	-0.22	-0.20	0.02	0.20	0.37	0.31	-0.03
Methanoculleus marisnigri JR1	0.62	-0.12	-0.06	0.02	0.68	0.22	-0.05	-0.04
Methanoregula boonei	0.27	-0.16	-0.23	0.09	0.44	0.23	0.12	-0.09
Methanospirillum hungatei	-0.15	-0.22	-0.24	0.00	0.05	0.31	0.29	-0.25
Methanocella paludicola	0.45	0.05	-0.02	0.11	0.64	0.04	-0.04	0.05
Methanosphaerula palustris	0.34	-0.12	-0.20	0.02	0.37	0.21	0.11	-0.14

Species	WCA of CU genome-wide				WCA of highly expressed CU			
	1	2	3	4	1	2	3	4
Euryarchaeaota								
Thermococcales								
Pyrococcus abyssi	-0.17	0.29	0.00	-0.14	-0.10	-0.26	0.12	0.14
Pyrococcus furiosus	-0.37	0.11	0.01	-0.06	-0.32	-0.13	0.12	0.12
Pyrococcus horikoshii	-0.30	0.18	-0.02	-0.08	-0.25	-0.23	0.15	0.05
Thermococcus gammatolerans	0.27	0.22	-0.06	-0.12	0.42	-0.12	0.19	0.21
Thermococcus kodakarensis	0.20	0.22	-0.08	-0.11	0.45	-0.11	0.28	0.25
Thermococcus onnurineus	0.19	0.20	-0.07	-0.13	0.39	-0.11	0.30	0.22
Thermococcus sibiricus	-0.36	0.01	-0.02	-0.01	-0.44	-0.02	0.08	0.02
Thermoplasmatales								
Picrophilus torridus	-0.39	0.09	-0.05	-0.01	-0.22	-0.24	0.17	-0.13
Thermoplasma acidophilum	-0.01	0.18	-0.13	-0.02	0.09	-0.25	0.10	-0.07
Thermoplasma volcanium	-0.30	0.09	-0.03	0.00	-0.24	-0.17	0.04	-0.16
Aciduliprofundum boonei	-0.39	0.06	0.01	0.03	-0.36	-0.12	-0.01	-0.14
Crenarchaeota								
Desulfurococcales								
Aeropyrum pernix	0.17	0.41	-0.03	0.01	0.29	-0.47	0.04	0.00
Hyperthermus butylicus	0.06	0.26	0.00	-0.09	0.19	-0.28	0.00	-0.10
Ignicoccus hospitalis	0.31	0.41	0.14	0.08	0.42	-0.41	-0.13	0.13
Staphylothermus marinus	-0.55	0.00	0.12	-0.06	-0.54	-0.03	-0.09	-0.14
Desulfurococcus kamchatkensis	-0.18	0.24	0.03	-0.05	-0.18	-0.26	-0.02	-0.08
Thermoproteales								
Caldivirga maquilingensis	-0.32	0.28	0.05	-0.17	-0.22	-0.39	0.13	-0.02
Pyrobaculum aerophilum	0.04	0.17	0.16	0.18	0.09	-0.17	-0.26	-0.02
Pyrobaculum arsenaticum	0.21	0.25	0.08	0.13	0.26	-0.21	-0.18	-0.04
Pyrobaculum calidifontis	0.27	0.28	0.11	0.18	0.27	-0.25	-0.20	-0.02
Pyrobaculum islandicum	-0.07	0.13	0.12	0.10	-0.21	-0.07	-0.23	-0.13
Thermoproteus neutrophilus	0.39	0.33	0.05	0.16	0.47	-0.28	-0.14	0.00
Thermofilum pendens	0.32	0.32	0.03	0.01	0.23	-0.24	-0.07	0.03
Sulfolobales								
Metallosphaera sedula	-0.10	0.26	-0.05	-0.02	-0.20	-0.23	0.04	-0.05
Sulfolobus acidocaldarius	-0.48	0.07	0.12	-0.02	-0.61	-0.10	-0.12	-0.08
Sulfolobus islandicus	-0.55	0.04	0.18	-0.02	-0.63	-0.09	-0.17	-0.09
Sulfolobus solfataricus	-0.50	0.06	0.17	-0.03	-0.63	-0.09	-0.17	-0.11
Sulfolobus tokodaii	-0.63	-0.05	0.17	0.00	-0.76	0.00	-0.19	-0.10
Thaumarchaeota								
Nitrosopumilus maritimus	-0.65	-0.25	0.04	0.01	-0.67	0.28	0.01	0.00
Cenarchaeum symbiosum A	0.29	0.13	-0.02	0.19	0.44	-0.19	-0.15	-0.12
Korarchaeota								
Korarchaeum cryptofilum	-0.03	0.40	-0.04	-0.05	0.08	-0.50	0.02	-0.05
Nanoarchaeota								
Nanoarchaeum equitans	-0.63	-0.07	0.31	0.14	-0.71	-0.03	-0.33	-0.02

Species	S[a]	GC3s[b]	Optimal codon[c]																	
			Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
Euryarchaeaota																				
Archaeoglobales																				
Archaeoglobus fulgidus	0.382*	0.56	AAC	GAC	--	--	UUC	UAC	AUC	CAG	GAG	AAG	--	GGU	CCG	--	GUC	AGA	CUC	UCA
Archaeoglobus profundus	0.65*	0.43	AAC	GAC	--	CAC	UUC	--	AUC	CAG	GAG	AAG	--	GGU	--	--	--	AGA	CUG	--
Ferroglobus placidus	0.716*	0.48	AAC	GAC	--	--	UUC	UAC	AUA	CAG	GAG	AAG	--	GGA	CCG	--	--	AGA	CUC	AGC
Halobacteriales																				
Haloarcula marismortui	1.032*	0.75	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCC	GGU	--	ACC	GUC	CGC	CUC	UCC
Halobacterium salinarum	0.959*	0.87	--	GAC	--	--	UUC	UAC	--	CAG	GAG	AAG	GCA	GGU	CCG	ACG	--	--	--	UCC
Halorubrum lacusprofundi	0.536*	0.86	AAC	GAC	--	--	--	UAC	--	CAG	--	AAG	GCC	GGC	--	ACG	GUC	--	CUG	UCC
Halomicrobium mukohataei	1.168*	0.85	AAC	GAC	--	--	UUC	UAC	AUC	CAG	GAA	AAG	GCA	--	CCC	ACC	GUC	CGU	CUG	UCC
Halorhabdus utahensis	1.43*	0.79	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCC	GGC	CCC	ACC	--	CGC	CUC	AGC
Haloterrigena turkmenica	0.925*	0.87	AAC	GAC	UGU	--	UUC	UAC	--	CAG	--	AAG	GCA	GGU	--	ACC	GUC	CGU	CUG	UCC
Haloquadratum walsbyi	0.080	0.42	--	--	--	--	UUC	--	AUU	--	--	AAG	GCU	--	--	--	GUU	--	CUU	--
Natrialba magadii	1.081*	0.76	AAC	GAC	UGU	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	--	ACC	GUC	CGU	CUG	UCC
Natronomonas pharaonis	1.254*	0.77	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	--	GGC	CCC	ACC	GUC	CGU	CUG	UCC
Methanobacteriales																				
Methanobacterium thermoautotrophicus	0.848*	0.54	AAC	GAC	--	CAC	UUC	UAC	AUC	--	GAA	AAG	GCA	GGU	CCA	ACA	GUC	CGA	CUC	UCA
Methanobrevibacter ruminantium	0.885*	0.30	AAC	GAC	UGU	CAC	UUC	UAC	AUC	CAA	GAA	AAA	GCU	GGU	CCU	ACC	GUA	AGA	UUA	UCC
Methanobrevibacter smithii	0.899*	0.18	AAC	GAC	UGU	CAC	--	UAC	AUC	CAA	GAA	AAA	GCU	GGU	CCU	ACA	GUA	AGG	CUC	UCU
Methanosphaera stadtmanae	0.976*	0.11	AAC	GAC	--	CAC	UUC	UAC	AUC	CAA	GAA	AAA	--	GGU	CCU	ACC	GUA	--	CUC	UCU
Methanopyrales																				
Methanopyrus kandleri	0.957*	0.74	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCC	GGU	CCG	ACC	GUG	--	CUG	AGC
Methanococcales																				
Methanocaldococcus jannaschii	1.138*	0.22	AAC	GAC	UGU	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCA	ACA	GUC	AGA	UUA	UCA
Methanocaldococcus fervens	1.178*	0.23	AAC	GAC	UGU	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCA	ACA	GUC	--	UUA	UCA
Methanocaldococcus vulcanius	0.489*	0.24	AAC	GAC	UGU	CAC	UUC	UAC	AUC	--	--	AAG	GCU	GGU	CCA	ACA	GUC	AGA	UUA	UCA
Methanococcus aeolicus	0.923*	0.20	AAC	GAC	--	CAC	UUC	UAC	AUC	--	GAA	--	GCU	GGU	CCU	ACA	GUU	AGA	CUU	UCA
Methanococcus maripaludis S2	1.805*	0.23	AAC	GAC	--	CAC	UUC	UAC	AUC	CAA	--	AAA	GCU	GGU	CCU	ACA	GUU	AGA	--	UCA
Methanococcus vannielii	1.461*	0.20	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAA	--	GCU	GGU	CCU	ACC	GUU	AGA	CUC	UCA

Species	S[a]	GC3s[b]	Optimal codon[c]																		
			Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
Euryarchaeaota																					
Methanosarcinales																					
Methanococcoides burtonii	0.910*	0.37	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCA	GGU	CCA	ACA	GUA	CGU	CUC	UCC
Methanosaeta thermophila PT	0.665*	0.64	AAC	--	--	CAC	UUC	--	--	CAG	--	AAG	GCC	GGC	--	--	GUC	AGG	--	--
Methanosarcina acetivorans	0.852*	0.44	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	--	ACC	GUC	CGU	CUC	UCC
Methanosarcina barkeri fusaro	0.671*	0.37	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	--	ACC	GUC	CGU	CUC	UCC
Methanosarcina mazei	0.892*	0.41	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCG	ACC	GUC	CGU	CUC	UCC
Rice Cluster I MRE50	1.058*	0.70	AAC	GAC	UGC	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCG	ACC	GUC	CGU	CUC	UCC
Methanomicrobiales																					
Methanocorpusculum labreanum	0.884*	0.56	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCG	ACC	GUC	CGU	CUU	UCC
Methanoculleus marisnigri JR1	0.470*	0.80	AAC	GAC	--	CAC	UUC	--	AUC	CAG	--	AAG	GCA	GGU	CCC	ACC	--	CGU	CUG	--
Methanoregula boonei	0.690*	0.64	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCC	ACC	GUC	CGU	CUC	--
Methanospirillum hungatei	0.628*	0.43	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGU	CCG	ACC	GUC	CGU	CUC	AGC
Methanocella paludicola	1.241*	0.74	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	--	AAG	GCA	GGC	CCC	ACC	GUC	AGA	CUC	UCC
Methanosphaerula palustris	0.335	0.67	AAC	--	--	--	UUC	--	AUC	CAG	--	AAG	GCA	GGU	CCA	--	GUC	CGU	CUG	--
Thermococcales																					
Pyrococcus abyssi	0.817*	0.49	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCU	GGU	CCA	ACC	--	AGA	CUC	AGC
Pyrococcus furiosus	0.620*	0.37	AAC	GAC	UGU	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCU	GGU	CCA	--	GUU	AGA	CUC	AGC
Pyrococcus horikoshii	0.474*	0.41	AAC	--	UGU	CAC	UUC	UAC	--	CAG	GAG	AAG	GCU	GGU	CCA	--	GUU	AGA	CUU	AGU
Thermococcus gammatolerans	0.679*	0.68	AAC	GAC	UGU	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCC	GGU	CCG	ACC	GUC	AGG	CUC	AGC
Thermococcus kodakarensis	1.541*	0.64	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCC	GGU	CCG	ACC	GUC	CGU	CUC	AGC
Thermococcus onnurineus	1.022*	0.63	AAC	GAC	--	CAC	UUC	UAC	AUC	CAG	GAG	AAG	GCU	GGU	CCG	ACC	GUC	CGU	CUC	AGC
Thermococcus sibiricus	0.079	0.36	AAC	--	--	CAC	--	--	AUU	--	--	AAG	GCU	GGU	CCA	--	GUU	AGA	CUU	UCA
Thermoplasmatales																					
Picrophilus torridus	0.607*	0.34	AAC	GAC	--	CAC	UUC	UAC	--	CAG	GAG	AAG	GCA	GGU	--	--	--	CGU	CUG	AGC
Thermoplasma acidophilum	0.451*	0.52	AAC	--	--	CAC	UUC	UAC	AUA	CAG	GAG	AAG	--	GGU	--	--	--	AGG	CUC	--
Thermoplasma volcanium	0.120	0.39	--	--	--	CAC	UUC	--	AUA	CAG	--	AAG	--	GGU	--	--	--	CGU	CUU	--
Aciduliprofundum boonei	0.142*	0.35	AAC	--	--	--	--	--	--	CAG	--	AAG	--	GGU	--	--	--	CGC	--	--

Species	S[a]	GC3s[b]	Optimal codon[c]																	
			Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
Crenarchaeota																				
Desulfurococcales																				
Aeropyrum pernix	0.356*	0.66	--	--	--	--	UUC	--	AUA	CAG	GAG	AAG	GCA	GGC	CCG	--	--	AGG	CUG	AGC
Hyperthermus butylicus	0.649*	0.57	AAC	GAC	--	--	UUC	UAC	AUU	CAG	GAG	AAG	GCA	GGC	CCA	--	GUA	CGU	CUA	AGC
Ignicoccus hospitalis	1.039*	0.75	AAC	GAC	--	--	UUC	UAC	AUC	CAG	--	AAG	--	GGC	CCC	ACC	--	AGG	CUG	AGC
Staphylothermus marinus	0.223*	0.26	AAC	--	--	--	UUC	UAC	--	--	--	--	--	GGA	--	--	GUA	CGU	CUA	AGC
Desulfurococcus kamchatkensis	0.006	0.47	--	--	--	--	--	--	AUA	--	--	--	--	GGU	CCA	--	GUA	--	--	--
Thermoproteales																				
Caldivirga maquilingensis	-0.013	0.43	--	--	--	--	--	--	AUA	CAG	GAG	AAG	--	GGU	--	--	GUU	AGG	CUU	UCU
Pyrobaculum aerophilum	0.077	0.58	--	--	--	--	--	--	AUC	--	--	AAG	--	GGC	--	--	GUC	--	--	UCA
Pyrobaculum arsenaticum	-0.024	0.66	--	--	--	--	--	--	--	CAG	GAG	--	--	GGU	CCC	--	--	--	--	UCG
Pyrobaculum calidifontis	-0.077	0.70	--	--	--	--	--	--	AUU	CAG	GAG	AAG	GCU	GGC	CCG	ACU	GUU	CGU	CUU	--
Pyrobaculum islandicum	-0.248	0.51	AAU	GAU	UGU	CAU	--	UAU	AUU	--	GAA	--	GCU	GGU	CCU	ACU	GUA	AGA	UUA	UCU
Thermoproteus neutrophilus	0.092	0.76	--	--	--	CAU	UUC	--	AUC	--	GAG	AAG	GCU	GGC	CCG	--	--	CGU	UUG	--
Thermofilum pendens	-0.275	0.72	--	GAU	--	--	UUU	UAU	AUA	--	GAA	--	GCA	GGA	--	ACU	GUU	AGA	--	UCU
Sulfolobales																				
Metallosphaera sedula	-0.133	0.51	--	--	UGU	--	--	--	--	--	--	--	--	GGU	CCA	ACU	GUA	AGA	--	--
Sulfolobus acidocaldarius	-0.184	0.31	AAU	GAU	--	--	--	UAU	AUA	--	--	--	GCA	GGU	--	ACU	GUA	AGA	UUA	--
Sulfolobus islandicus	-0.058	0.28	--	GAU	UGU	--	--	UAU	AUA	CAG	GAA	--	--	--	--	--	--	AGA	UUA	--
Sulfolobus solfataricus	-0.258	0.31	--	GAU	--	CAU	--	UAU	AUA	--	GAA	AAA	GCA	GGU	CCA	ACU	GUA	AGA	UUA	--
Sulfolobus tokodaii	-0.105	0.23	AAU	--	--	--	--	UAU	AUA	--	GGU	--	--	--	CCA	ACA	GUA	AGA	UUA	AGC
Thaumarchaeota																				
Nitrosopumilus maritimus	0.745*	0.20	AAC	GAC	--	CAC	UUC	UAC	AUC	--	--	AAG	GCA	GGU	CCA	--	GUC	AGA	CUC	AGU
Cenarchaeum symbiosum A	0.532*	0.68	AAC	GAC	--	--	UUC	--	AUA	CAG	GAA	AAG	--	GGC	--	--	GUC	AGG	CUC	UCG
Korarchaeota																				
Korarchaeum cryptofilum	0.181	0.57	AAC	--	--	--	--	--	AUA	CAG	GAG	AAG	--	GGU	--	ACG	--	AGG	CUG	AGC
Nanoarchaeota																				
Nanoarchaeum equitans	0.185*	0.23	--	--	--	--	UUC	--	--	--	--	--	GCA	GGU	CCA	--	--	--	UUA	UCU

Code	tRNA genes																																										
	Asn	**Asp**	**Cys**	**His**	**Phe**	**Tyr**	**Ile**		**Gln**		**Glu**		**Lys**		**Ala**			**Gly**			**Pro**			**Thr**			**Val**			**Arg**					**Leu**					**Ser**			
	GTT	**GTC**	**GCA**	**GTG**	**GAA**	**GTA**	**GAT**	**TAT**	**CTG**	**TTG**	**TTC**	**CTC**	**TTT**	**CTT**	**GGC**	**TGC**	**CGC**	**GCC**	**TCC**	**CCC**	**GGG**	**TGG**	**CGG**	**GGT**	**TGT**	**CGT**	**GAC**	**TAC**	**CAC**	**GCG**	**TCG**	**CCG**	**TCT**	**CCT**	**GAG**	**TAG**	**CAG**	**TAA**	**CAA**	**GGA**	**TGA**	**CGA**	**GCT**
Euryarchaeaota																																											
Archaeoglobales																																											
Arcful	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Arcpro	1	1	1	1	1	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Ferpla	1	1	1	1	1	1	1	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	1	1	2	1	1	1	1	1	1	1	1
Halobacteriales																																											
Halmar	1	2	1	1	1	1	1	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Halsal	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hallac	1	2	1	1	2	1	1	0	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Halmuk	1	2	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
Haluta	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Haltur	1	2	1	1	1	1	1	0	0	1	1	1	1	1	1	3	1	2	1	1	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
Halwal	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Natmag	1	2	1	1	1	1	1	0	0	1	1	1	1	1	1	2	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Natpha	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Methanobacteriales																																											
Metthe	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	2	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	2	1	0	1
Metrum	2	1	1	2	2	2	2	0	0	1	2	0	2	1	1	3	0	1	2	0	0	2	0	1	1	0	2	2	0	1	1	0	1	1	1	0	0	2	1	2	2	0	2
Metsmi	1	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1	0	1	1	0	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	0	1	2	0	2
Metsta	2	1	1	1	2	1	1	0	1	1	2	0	1	0	1	3	0	1	2	0	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	0	2	0	1	1	0	1
Methanopyrales																																											
Metkan	1	1	1	1	1	1	1	0	0	1	2	0	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	0	1	1	0	1
Methanococcales																																											
Metjan	1	1	1	1	1	1	1	0	0	1	2	0	1	0	1	2	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	0	1
Metfer	1	1	1	1	1	1	1	0	0	1	2	0	1	0	1	1	2	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	3	0	1	0	1	1	0	1
Metvul	1	1	1	1	1	1	1	0	0	1	2	0	1	0	1	2	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	0	1
Metaeo	1	2	1	1	1	1	1	0	0	1	2	0	2	0	0	0	0	1	2	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	0	1	1	0	1
MetmarS2	1	2	1	1	1	1	1	0	0	1	2	0	2	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	0	1	1	0	1
Metvan	1	2	1	1	1	1	1	0	0	1	2	0	2	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	0	1	1	0	1

Code	tRNA genes																																										
	Asn	**Asp**	**Cys**	**His**	**Phe**	**Tyr**	**Ile**		**Gln**		**Glu**		**Lys**		**Ala**			**Gly**			**Pro**			**Thr**			**Val**			**Arg**					**Leu**					**Ser**			
	GTT	**GTC**	**GCA**	**GTG**	**GAA**	**GTA**	**GAT**	**TAT**	**CTG**	**TTG**	**TTC**	**CTC**	**TTT**	**CTT**	**GGC**	**TGC**	**CGC**	**GCC**	**TCC**	**CCC**	**GGG**	**TGG**	**CGG**	**GGT**	**TGT**	**CGT**	**GAC**	**TAC**	**CAC**	**GCG**	**TCG**	**CCG**	**TCT**	**CCT**	**GAG**	**TAG**	**CAG**	**TAA**	**CAA**	**GGA**	**TGA**	**CGA**	**GCT**
Euryarchaeaota																																											
Methanosarcinales																																											
Metbur	1	1	3	1	1	1	1	0	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	0	1
MetthePT	1	1	2	1	1	1	1	0	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
Metace	2	1	3	1	2	1	2	0	1	2	2	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1
Metbak	2	1	4	1	2	1	2	0	1	1	2	1	1	1	1	1	1	3	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	2	1	1	4	1	1	1	1	1
Metmaz	2	1	3	1	2	1	2	0	1	1	2	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
RicClu	1	1	5	1	1	1	2	0	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Methanomicrobiales																																											
Metlab	1	2	4	1	1	1	1	0	1	1	1	1	1	1	1	3	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MetmarJR	1	1	1	1	1	1	1	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
Metboo	1	1	4	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Methun	1	2	2	1	1	1	1	0	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Metpa2	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	2	1	1	1	1	0	0	1	2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1
Metpal	1	1	8	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Thermococcales																																											
Pyraby	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pyrfur	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pyrhor	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Thegam	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Thekod	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Theonu	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
Thesib	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Thermoplasmatales																																											
Pictor	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Theaci	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
Thevol	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
Aciboo	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1

Code	tRNA genes																																										
	Asn	**Asp**	**Cys**	**His**	**Phe**	**Tyr**	**Ile**		**Gln**		**Glu**		**Lys**		**Ala**			**Gly**			**Pro**			**Thr**			**Val**			**Arg**					**Leu**					**Ser**			
	GTT	**GTC**	**GCA**	**GTG**	**GAA**	**GTA**	**GAT**	**TAT**	**CTG**	**TTG**	**TTC**	**CTC**	**TTT**	**CTT**	**GGC**	**TGC**	**CGC**	**GCC**	**TCC**	**CCC**	**GGG**	**TGG**	**CGG**	**GGT**	**TGT**	**CGT**	**GAC**	**TAC**	**CAC**	**GCG**	**TCG**	**CCG**	**TCT**	**CCT**	**GAG**	**TAG**	**CAG**	**TAA**	**CAA**	**GGA**	**TGA**	**CGA**	**GCT**
Crenarchaeota																																											
Desulfurococcales																																											
Aerper	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	3	1	1	1	1	1	1	1	1	1	1	1
Hypbut	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1
Ignhos	1	1	1	1	1	1	1	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Stamar	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
Deskam	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	2	1	1	1	1	1	1	1	1	1
Thermoproteales																																											
Calmaq	1	1	1	1	1	1	1	0	1	1	2	1	1	1	1	1	1	3	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1
Pyraer	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pyrars	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pyrcal	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pyrisl	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Theneu	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Thepen	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sulfolobales																																											
Metsed	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sulaci	1	2	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1
Sulisl	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
Sulsol	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sultok	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Thaumarchaeota																																											
Nitmar	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Censym	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Korarchaeota																																											
Korcry	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Nanoarchaeota																																											
Natequ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1

Species	*S*	*S* adapted for each amino acid group																	
		Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
Euryarchaeaota																			
Archaeoglobales																			
Archaeoglobus fulgidus	0.382*	0.46	0.36	0.22	0.40	0.70	0.44	0.15	1.13	0.26	0.71	0.06	0.43	0.48	0.17	0.31	0.55	0.43	0.34
Archaeoglobus profundus	0.650*	0.84	0.70	0.09	0.69	0.94	0.33	0.63	1.70	0.32	0.83	0.14	0.48	0.23	0.13	0.22	0.71	0.44	0.35
Ferroglobus placidus	0.716*	1.06	0.63	0.23	0.47	1.16	0.57	0.22	0.59	0.23	0.60	0.20	0.36	0.25	0.26	0.21	0.47	0.41	0.55
Halobacteriales																			
Haloarcula marismortui	1.032*	0.77	0.90	0.32	1.81	2.16	1.94	0.34	0.99	0.14	0.70	0.49	0.47	0.20	0.51	0.43	0.65	0.21	0.80
Halobacterium salinarum	0.959*	0.62	0.32	0.48	0.15	2.29	2.09	0.16	0.57	0.35	0.63	0.46	0.30	0.25	0.23	0.33	0.28	0.11	0.62
Halorubrum lacusprofundi	0.536*	0.97	0.52	0.00	0.35	0.77	0.99	0.09	0.71	0.11	0.63	0.32	0.44	0.14	0.24	0.21	0.65	0.29	0.81
Halomicrobium mukohataei	1.168*	1.38	0.90	0.33	0.84	2.19	1.13	0.56	0.95	0.17	0.81	0.72	0.27	0.25	0.44	0.37	0.78	0.27	0.74
Halorhabdus utahensis	1.430*	1.01	0.96	0.45	0.63	3.23	0.65	1.18	1.42	0.34	0.98	0.40	0.35	0.29	0.24	0.15	0.41	0.38	0.29
Haloterrigena turkmenica	0.925*	1.05	0.81	1.03	0.87	2.68	1.24	0.09	0.60	0.09	0.38	0.71	0.61	0.20	0.28	0.22	1.07	0.21	0.52
Haloquadratum walsbyi	0.080	0.06	0.01	0.32	0.11	0.52	0.06	0.09	0.02	0.04	0.42	0.22	0.17	0.11	0.04	0.17	0.34	0.34	0.23
Natrialba magadii	1.081*	2.09	0.90	1.29	1.38	1.19	1.33	0.30	1.20	0.07	1.00	0.32	0.38	0.27	0.30	0.47	0.94	0.42	0.63
Natronomonas pharaonis	1.254*	2.33	0.97	0.33	0.98	1.91	0.82	0.31	0.67	0.06	0.64	0.19	0.29	0.54	0.30	0.47	0.48	0.32	0.70
Methanobacteriales																			
Methanobacterium thermoautotrophicus	0.848*	1.28	0.95	0.06	1.04	0.64	0.79	0.57	0.42	0.21	0.35	0.26	0.53	0.30	0.28	0.44	0.60	0.43	0.51
Methanobrevibacter ruminantium	0.885*	1.37	0.85	1.32	2.24	0.89	1.20	0.43	1.51	1.23	0.76	0.80	1.18	0.35	1.16	1.00	0.72	0.94	0.63
Methanobrevibacter smithii	0.899*	1.08	0.29	0.86	1.56	1.24	0.83	0.52	1.39	0.11	0.86	0.24	0.79	0.43	0.70	0.32	0.35	0.81	0.63
Methanosphaera stadtmanae	0.976*	1.29	0.50	0.96	1.88	1.15	1.03	0.47	0.59	0.58	1.38	0.09	0.23	0.40	0.38	0.47	0.41	0.59	0.68
Methanopyrales																			
Methanopyrus kandleri	0.957*	1.20	1.22	0.26	0.59	2.11	0.82	0.49	1.39	0.55	1.48	0.18	0.79	0.52	0.23	0.18	0.10	0.70	0.72
Methanococcales																			
Methanocaldococcus jannaschii	1.138*	1.60	0.79	0.82	1.75	1.45	1.23	0.55	0.55	0.06	0.33	0.42	0.39	0.93	0.76	0.65	1.00	0.51	0.91
Methanocaldococcus fervens	1.178*	1.70	0.77	0.95	1.44	1.43	1.30	0.69	0.37	0.10	0.23	0.26	0.35	0.75	0.91	0.65	1.17	0.51	0.84
Methanocaldococcus vulcanius	0.489*	1.03	0.42	1.03	1.01	0.63	0.38	0.11	0.17	0.07	0.28	0.25	0.25	0.51	0.46	0.60	0.65	0.32	0.60
Methanococcus aeolicus	0.923*	1.11	0.47	0.26	1.03	1.26	1.22	0.37	0.20	0.25	0.02	0.67	0.47	0.56	0.36	0.37	0.64	0.26	0.33
Methanococcus maripaludis	1.805*	2.31	0.80	0.10	1.78	2.01	2.13	1.02	0.41	0.04	0.50	1.00	0.83	0.38	0.56	0.41	1.39	0.57	0.60
Methanococcus vannielii	1.461*	1.57	0.83	0.19	1.49	1.78	1.54	1.02	0.20	0.47	0.03	0.66	0.79	0.31	0.36	0.25	1.17	0.72	0.45

Species	*S*	*S* adapted for each amino acid group																	
		Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
Euryarchaeaota																			
Methanosarcinales																			
Methanococcoides burtonii	0.910*	1.10	0.27	0.02	1.62	1.16	0.73	0.61	1.83	0.37	0.76	0.51	0.70	0.53	0.27	0.35	0.54	0.65	0.48
Methanosaeta thermophila PT	0.665*	0.54	0.16	0.22	0.40	0.84	0.23	0.66	1.32	0.26	0.95	0.27	0.19	0.16	0.11	0.43	0.22	0.09	0.09
Methanosarcina acetivorans	0.852*	1.20	1.16	0.33	1.73	1.05	0.75	0.47	1.83	0.19	1.01	0.62	0.61	0.22	0.64	0.28	0.48	0.47	0.78
Methanosarcina barkeri	0.671*	0.82	0.64	0.26	1.24	1.05	0.53	0.39	1.85	0.05	1.00	0.37	0.38	0.11	0.67	0.29	0.60	0.41	0.51
Methanosarcina mazei	0.892*	1.23	0.86	0.57	1.89	1.10	0.74	0.58	2.02	0.08	1.16	0.51	0.73	0.38	0.67	0.22	0.54	0.46	0.61
Rice Cluster I MRE50	1.058*	1.29	1.01	1.30	1.85	1.53	1.28	0.58	1.11	0.05	1.60	0.79	0.72	0.29	0.80	0.63	0.98	0.69	0.76
Methanomicrobiales																			
Methanocorpusculum labreanum	0.884*	1.29	0.51	0.40	2.07	1.34	1.28	0.36	2.60	0.05	0.97	1.21	0.99	0.42	0.97	0.32	1.00	0.39	0.97
Methanoculleus marisnigri JR1	0.470*	0.47	0.33	0.04	0.81	0.72	0.33	0.54	2.34	0.01	0.85	0.53	0.83	0.28	0.33	0.12	0.86	0.42	0.15
Methanoregula boonei	0.690*	1.00	0.43	0.47	1.10	1.07	0.76	0.18	1.74	0.09	1.27	0.25	0.58	0.25	0.45	0.57	1.12	0.45	0.17
Methanospirillum hungatei	0.628*	0.98	0.54	0.35	1.24	0.69	0.55	0.36	2.25	0.12	1.04	0.50	0.69	0.51	0.52	0.51	1.41	0.31	0.51
Methanocella paludicola	1.241*	1.59	0.65	0.18	1.80	1.20	1.39	1.00	2.59	0.01	1.50	0.34	0.93	0.28	0.92	0.61	0.62	0.35	0.39
Methanosphaerula palustris	0.335	0.49	0.14	0.29	0.34	0.58	0.26	0.15	0.98	0.02	0.94	0.69	0.49	0.63	0.28	0.31	0.75	0.19	0.28
Thermococcales																			
Pyrococcus abyssi	0.817*	1.16	0.47	0.34	0.73	0.44	0.58	0.11	0.54	0.22	0.86	0.26	0.91	0.55	0.34	0.20	0.30	0.25	0.58
Pyrococcus furiosus	0.620*	0.90	0.27	1.04	1.28	0.92	0.40	0.31	0.66	0.26	1.05	0.39	0.87	0.47	0.09	0.20	0.59	0.33	0.40
Pyrococcus horikoshii	0.474*	0.59	0.21	1.08	0.71	0.70	0.33	0.10	0.75	0.22	0.94	0.31	0.78	0.26	0.12	0.37	0.27	0.39	0.47
Thermococcus gammatolerans	0.679*	1.30	0.66	0.92	1.05	0.70	0.65	0.32	1.55	0.51	1.13	0.35	0.91	0.54	0.99	0.61	0.95	0.42	0.65
Thermococcus kodakarensis	1.541*	3.11	0.99	0.35	2.20	1.55	1.12	0.79	2.94	0.85	2.15	0.50	1.24	0.90	1.33	0.79	1.04	1.11	0.59
Thermococcus onnurineus	1.022*	1.84	0.62	0.59	2.01	0.92	0.74	0.69	1.43	0.74	1.73	0.45	0.84	0.99	1.47	0.84	0.89	0.85	0.39
Thermococcus sibiricus	0.079	0.47	0.03	0.10	0.34	0.18	0.00	0.19	0.18	0.03	0.42	0.31	0.77	0.63	0.21	0.29	0.56	0.57	0.42
Thermoplasmatales																			
Picrophilus torridus	0.607*	0.79	0.28	0.22	0.54	0.82	0.44	0.29	0.90	0.66	1.26	0.20	0.23	0.19	0.11	0.23	0.88	0.50	0.29
Thermoplasma acidophilum	0.451*	0.60	0.12	0.09	0.55	0.48	0.53	0.08	0.85	0.54	0.33	0.12	0.43	0.18	0.11	0.06	0.36	0.25	0.13
Thermoplasma volcanium	0.120	0.17	0.11	0.42	0.65	0.28	0.04	0.06	0.49	0.14	0.44	0.10	0.27	0.07	0.16	0.14	0.52	0.22	0.13
Aciduliprofundum boonei	0.142*	0.31	0.02	0.06	0.01	0.28	0.06	0.06	0.48	0.02	0.36	0.13	0.43	0.23	0.13	0.16	0.45	0.20	0.22

Species	*S*	*S* adapted for each amino acid group																	
		Asn	Asp	Cys	His	Phe	Tyr	Ile	Gln	Glu	Lys	Ala	Gly	Pro	Thr	Val	Arg	Leu	Ser
Crenarchaeota																			
Desulfurococcales																			
Aeropyrum pernix	0.356*	0.34	0.23	0.02	0.19	0.72	0.01	0.20	1.18	0.29	1.01	0.27	0.27	0.42	0.17	0.16	0.58	0.23	0.58
Hyperthermus butylicus	0.649*	0.94	0.47	0.37	0.51	0.82	0.47	0.50	0.73	0.46	1.14	0.27	0.59	0.45	0.13	0.31	0.43	0.48	0.44
Ignicoccus hospitalis	1.039*	1.62	0.65	0.06	0.72	1.14	0.86	0.74	0.38	0.15	0.85	0.13	0.55	0.22	0.51	0.12	0.48	0.24	0.79
Staphylothermus marinus	0.223*	0.47	0.02	0.02	0.01	0.37	0.28	0.04	0.25	0.16	0.12	0.18	0.31	0.13	0.22	0.21	0.57	0.24	0.52
Desulfurococcus kamchatkensis	0.006	0.18	0.10	0.38	0.06	0.23	0.10	0.17	0.25	0.02	0.18	0.11	0.43	0.35	0.24	0.22	0.25	0.13	0.23
Thermoproteales																			
Caldivirga maquilingensis	-0.013	0.24	0.13	0.54	0.01	0.24	0.13	0.66	0.35	0.45	0.89	0.13	0.29	0.18	0.18	0.30	0.62	0.44	0.48
Pyrobaculum aerophilum	0.077	0.16	0.18	0.45	0.44	0.23	0.15	0.15	0.19	0.14	0.37	0.09	0.29	0.11	0.13	0.29	0.11	0.14	0.30
Pyrobaculum arsenaticum	-0.024	0.25	0.03	0.28	0.17	0.18	0.08	0.28	0.31	0.38	0.19	0.16	0.28	0.20	0.15	0.12	0.28	0.12	0.34
Pyrobaculum calidifontis	-0.077	0.15	0.14	0.04	0.20	0.03	0.13	0.24	0.50	0.25	0.40	0.41	0.18	0.29	0.38	0.27	0.46	0.33	0.22
Pyrobaculum islandicum	-0.248	0.36	0.45	0.86	0.60	0.19	0.38	0.25	0.09	0.36	0.08	0.32	0.28	0.32	0.28	0.47	0.44	0.53	0.32
Thermoproteus neutrophilus	0.092	0.48	0.21	0.04	0.47	0.35	0.17	0.01	0.28	0.54	0.58	0.28	0.50	0.38	0.28	0.18	0.72	0.38	0.22
Thermofilum pendens	-0.275	0.18	0.40	0.06	0.20	0.47	0.58	0.23	0.32	0.36	0.12	0.48	0.55	0.19	0.36	0.36	0.45	0.32	0.42
Sulfolobales																			
Metallosphaera sedula	-0.133	0.32	0.16	0.40	0.35	0.02	0.38	0.16	0.02	0.03	0.23	0.35	0.49	0.52	0.31	0.43	0.52	0.39	0.31
Sulfolobus acidocaldarius	-0.184	0.29	0.62	0.45	0.12	0.22	0.39	0.45	0.10	0.14	0.11	0.36	0.23	0.10	0.23	0.25	0.42	0.48	0.18
Sulfolobus islandicus	-0.058	0.14	0.48	2.32	0.13	0.17	0.48	0.05	0.36	0.26	0.01	0.14	0.15	0.12	0.17	0.15	0.27	0.34	0.24
Sulfolobus solfataricus	-0.258	0.02	0.39	0.12	0.44	0.04	0.59	0.44	0.26	0.40	0.18	0.25	0.28	0.38	0.24	0.25	0.45	0.38	0.15
Sulfolobus tokodaii	-0.105	0.33	0.21	0.21	0.26	0.22	0.41	0.21	0.12	0.48	0.17	0.08	0.36	0.33	0.27	0.36	0.75	0.31	0.40
Thaumarchaeota																			
Nitrosopumilus maritimus	0.745*	1.11	0.40	0.05	1.15	1.14	0.47	0.41	0.33	0.10	0.23	0.30	0.56	0.45	0.12	0.42	0.77	0.50	0.28
Cenarchaeum symbiosum A	0.532*	1.03	0.59	0.75	0.20	0.33	0.26	0.69	0.66	0.17	0.79	0.05	0.45	0.04	0.13	0.45	0.21	0.38	0.22
Korarchaeota																			
Korarchaeum cryptofilum	0.181	0.27	0.07	0.11	0.15	0.36	0.03	0.44	0.92	0.27	0.69	0.13	0.22	0.21	0.34	0.12	0.41	0.21	0.31
Nanoarchaeota																			
Nanoarchaeum equitans	0.185*	0.21	0.14	0.22	0.07	0.45	0.15	0.18	0.82	0.12	0.02	0.21	0.30	0.47	0.08	0.14	0.22	0.17	0.45

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 1203–1212
doi:10.1098/rstb.2009.0305

Forces that influence the evolution of codon bias

Paul M. Sharp*, Laura R. Emery and Kai Zeng

Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, Edinburgh EH9 3JT, UK

The frequencies of alternative synonymous codons vary both among species and among genes from the same genome. These patterns have been inferred to reflect the action of natural selection. Here we evaluate this in bacteria. While intragenomic variation in many species is consistent with selection favouring translationally optimal codons, much of the variation among species appears to be due to biased patterns of mutation. The strength of selection on codon usage can be estimated by two different approaches. First, the extent of bias in favour of translationally optimal codons in highly expressed genes, compared to that in genes where selection is weak, reveals the long-term effectiveness of selection. Here we show that the strength of selected codon usage bias is highly correlated with bacterial growth rate, suggesting that selection has favoured translational efficiency. Second, the pattern of bias towards optimal codons at polymorphic sites reveals the ongoing action of selection. Using this approach we obtained results that were completely consistent with the first method; importantly, the frequency spectra of optimal codons at polymorphic sites were similar to those predicted under an equilibrium model. Highly expressed genes in *Escherichia coli* appear to be under continuing strong selection, whereas selection is very weak in genes expressed at low levels.

Keywords: codon usage; bacteria; population genetics; selection; mutation bias

1. INTRODUCTION

When the genetic code was decrypted in the 1960s, it became apparent that most amino acids are encoded by multiple (two to six) codons, which typically differ only at the third nucleotide of the codon. With the introduction of DNA sequencing in the late 1970s, it emerged that these alternative synonymous codons are not used with equal frequencies. Two phenomena were soon apparent: patterns of codon usage vary among species (Grantham *et al.* 1980), and in the model bacterium *Escherichia coli* (for which most data were available), codon usage is more biased in genes expressed at higher levels (Post & Nomura 1980; Gouy & Gautier 1982; see table 1). Both phenomena were interpreted as reflecting the action of natural selection.

The selective differences among synonymous codons reflect two aspects of the transfer RNA (tRNA) population present in the cell (Ikemura 1985). First, for some amino acids there are multiple species of tRNAs with different anticodons, and it is those codons translated by the most abundant tRNA species which are preferred in highly expressed genes. For example, there are five different Leu tRNAs in *E. coli*, but that with anticodon CAG is much more abundant than the others. This anticodon is complementary to the codon CUG, which is used nearly 20 times more often than any of the other five Leu codons in highly expressed genes (table 1). Second, many tRNAs can translate more than one codon, but with variable ability; the codon best recognized by the anticodon is preferred in highly expressed genes. For example, there is a single Phe tRNA in *E. coli*, with anticodon GAA, which translates both UUU and UUC; however, UUC is perfectly complementary to the anticodon, and is used about three times more often than UUU in highly expressed genes (table 1). Thus, from knowledge of the tRNA population it is possible to predict which codons are translationally optimal; namely, those that are best recognized by the most abundant tRNA species.

There has been much debate about exactly why translationally optimal codons are selected. The traditional view is that use of optimal codons increases the efficiency of translation (Ehrenberg & Kurland 1984; Andersson & Kurland 1990). Ribosomes constitute about two-thirds of the protein content of an *E. coli* cell when growing rapidly (Pedersen *et al.* 1978), and the abundance of ribosomes may be the main factor limiting growth rate. Optimal codons may be translated faster than non-optimal codons (Sørensen & Pedersen 1991), such that ribosomes move faster along an mRNA containing more optimal codons, and the ribosomes are more quickly released to be available to translate other mRNAs. Thus, use of optimal codons, especially in genes expressed at high levels encoding mRNAs that must be translated more often, allows more efficient use of ribosomes and leads to faster growth rate (Kudla *et al.* 2009), conferring an obvious selective advantage, at least in bacteria occupying certain niches.

An alternative view is that the use of optimal codons increases the accuracy of translation. Sites where the identity of the amino acid is more critical for protein

* Author for correspondence (paul.sharp@ed.ac.uk).

One contribution of 16 to a Theme Issue 'The population genetics of mutations: good, bad and indifferent' dedicated to Brian Charlesworth on his 65th birthday.

Table 1. Codon usage in *E. coli*. Codon usage is compared between a set of 40 highly expressed genes (high; see Sharp *et al.* 2005) and the genome as a whole (all); the data are relative synonymous codon usage values (the ratio of the observed number to that expected if all codons for an amino acid were used equally). Nineteen codons occurring at significantly higher frequencies (see Henry & Sharp 2007) in the high dataset are shown in bold. The data are for *E. coli* strain K-12 MG1655 (accession number U00096).

		high	all			high	all			high	all			high	all
Phe	UUU	0.45	1.15	Ser	**UCU**	2.54	0.87	Tyr	UAU	0.48	1.14	Cys	UGU	0.77	0.89
Phe	**UUC**	1.55	0.85	Ser	**UCC**	1.52	0.89	Tyr	**UAC**	1.52	0.86	Cys	UGC	1.23	1.11
Leu	UUA	0.14	0.79	Ser	UCA	0.19	0.74	Ter	UAA	2.85	1.89	Ter	UGA	0.15	0.88
Leu	UUG	0.25	0.77	Ser	UCG	0.06	0.92	Ter	UAG	0	0.23	Trp	UGG	1.00	1.00
Leu	CUU	0.30	0.62	Pro	CCU	0.59	0.64	His	CAU	0.61	1.14	Arg	**CGU**	4.13	2.27
Leu	CUC	0.23	0.63	Pro	CCC	0.05	0.50	His	**CAC**	1.39	0.86	Arg	CGC	1.80	2.39
Leu	CUA	0.01	0.22	Pro	CCA	0.52	0.76	Gln	CAA	0.37	0.69	Arg	CGA	0	0.39
Leu	**CUG**	5.07	2.97	Pro	**CCG**	2.85	2.10	Gln	**CAG**	1.63	1.31	Arg	CGG	0.03	0.59
Ile	AUU	0.72	1.52	Thr	**ACU**	1.89	0.67	Asn	AAU	0.25	0.90	Ser	AGU	0.29	0.91
Ile	**AUC**	2.27	1.26	Thr	ACC	1.75	1.74	Asn	**AAC**	1.75	1.10	Ser	AGC	1.40	1.66
Ile	AUA	0.01	0.22	Thr	ACA	0.19	0.53	Lys	AAA	1.45	1.53	Arg	AGA	0.03	0.23
Met	AUG	1.00	1.00	Thr	ACG	0.17	1.07	Lys	AAG	0.55	0.47	Arg	AGG	0	0.14
Val	**GUU**	2.07	1.03	Ala	**GCU**	1.83	0.65	Asp	GAU	0.69	1.26	Gly	**GGU**	2.53	1.35
Val	GUC	0.30	0.86	Ala	GCC	0.32	1.08	Asp	**GAC**	1.31	0.74	Gly	GGC	1.39	1.61
Val	**GUA**	1.14	0.62	Ala	**GCA**	1.09	0.85	Glu	**GAA**	1.51	1.38	Gly	GGA	0.03	0.44
Val	GUG	0.48	1.49	Ala	GCG	0.76	1.42	Glu	GAG	0.49	0.62	Gly	GGG	0.06	0.60

function are expected to be more conserved across species, and also expected to be the sites where accuracy of translation is more important. The fruitfly, *Drosophila melanogaster*, exhibits stronger codon usage bias in more highly expressed genes, analogous to the situation in *E. coli* (Shields *et al.* 1988; Duret & Mouchiroud 1999), and it was found that codons for conserved amino acids have stronger codon bias in *D. melanogaster* (Akashi 1994). This accuracy hypothesis has the potential to explain the observation, otherwise surprising, that rates of non-synonymous and synonymous nucleotide substitution are correlated across genes in comparisons between *E. coli* and its close relative *Salmonella enterica* (Sharp 1991). Based on a variety of observations, some authors have concluded that translational accuracy is the primary object of codon selection in *E. coli* (Stoletzki & Eyre-Walker 2006), and indeed the dominant constraint on gene sequence evolution across both bacteria and eukaryotes (Drummond & Wilke 2008).

In this article we will focus on bacteria, examining the extent to which natural selection is responsible for the variations in codon usage seen among species and within genomes. In particular, we will contrast the results of two different approaches to estimating the strength of selection on codon usage bias. Finally, we will discuss the implications of the results, including their relevance to the efficiency versus accuracy debate introduced above.

2. VARIATION IN CODON USAGE BIAS AMONG BACTERIA

Analyses of bacteria other than *E. coli* have revealed that codon usage patterns vary among species in a number of ways. Most of the differences appear to be due, ultimately, to variations in mutation biases. First, it had been known for half a century that base composition, summarized by G + C content in double-stranded DNA, varies greatly among bacteria (Belozersky & Spirin 1958). Among published bacterial genome sequences, values of G + C content range from 17 per cent in *Carsonella ruddii* (Nakabachi *et al.* 2006) to 73 per cent in *Frankia alni* (Normand *et al.* 2007). This variation has long been viewed as the primary influence on codon usage differences between species of bacteria (Bibb *et al.* 1984; Muto & Osawa 1987). This has been confirmed by multivariate analyses comparing total genomic codon usage among bacteria, which showed that the single most important source of variation is G + C content (Lynn *et al.* 2002; Chen *et al.* 2004). It has often been speculated that this variation reflects the action of selection. In particular, it has been suggested that there would be pressure on thermophilic bacteria to have more G + C-rich genomes, because they are more thermostable (Bernardi & Bernardi 1986; Musto *et al.* 2004). However, most analyses have failed to find any correlation between growth temperature and genomic G + C content (e.g. Galtier & Lobry 1997; Lynn *et al.* 2002). Overall, the variation in G + C content is most simply explained by subtle but persistent mutation biases (Sueoka 1962).

Second, genome sequencing has revealed that in many bacteria base composition varies systematically between the leading and lagging strands of replication, with the leading strand being more G + T-rich (Lobry 1996; McLean *et al.* 1998). This strand-specific bias impacts codon usage, but the strength of the effect varies considerably among species. In the spirochaetes *Borrelia burgdorferi* (the cause of Lyme disease) and *Treponema denticola* (the cause of syphilis), strand-specific bias dominates codon usage variation among genes (Lafay *et al.* 1999); in other species the effect is much weaker, or undetectable (Kloster & Tang 2008). The source of this strand-specific bias has been debated, but the predominant ideas concern mutation biases. The leading and lagging strands are

replicated by different mechanisms with different mutation rates (Fijalkowska *et al.* 1998), which could lead to the observed differences in base composition. Alternatively, since there is an excess of genes located on the leading strand in many bacteria (Brewer 1988; Tillier & Collins 2000), biases in transcription-coupled repair could lead to a skew between the strands in nucleotide composition (Francino *et al.* 1996).

Third, for some amino acids, the identity of the translationally optimal codon varies among species. For example, in *Clostridium perfringens*, the codons heavily used in highly expressed genes (Musto *et al.* 2003) differ from those in *E. coli* (table 1) for six amino acids. These differences are correlated with changes in tRNA populations. While tRNA abundances have been measured for very few species, it is known that tRNA abundance is correlated with tRNA gene copy number (Kanaya *et al.* 1999), and so the latter may be used to predict the most abundant tRNAs. In the *E. coli* genome, where there are eight genes encoding five different Leu tRNAs, four genes encode the tRNA with the CAG anticodon (mentioned above as being the most abundant Leu tRNA species in *E. coli*). The *C. perfringens* genome also contains eight Leu tRNA genes (for four different tRNAs), but four encode the tRNA with anticodon UAA; the heavily used Leu codon is UUA, perfectly complementary to this predicted most abundant tRNA. Thus, there is co-adaptation between the codon usage of highly expressed genes and the tRNA population in both species, but the identity of the co-adapted state differs. Exactly how this divergence can occur is unclear, but it has been hypothesized that it could be driven by pressure from biased mutation patterns (Shields 1990).

Fourth, not all bacterial species exhibit the same clear trend in codon usage patterns associated with gene expression level. For example, in *Helicobacter pylori* (a bacterium that causes stomach ulcers), there is at most a very minor difference in codon usage between highly expressed and other genes (Lafay *et al.* 2000; Kloster & Tang 2008), while in *B. burgdorferi* most of the highly expressed genes are located on the leading strand of replication, and have G + T-rich codon usage that does not differ from other genes on that strand (Lafay *et al.* 1999). This difference among species most likely reflects variation in the extent to which natural selection is effective in shaping codon usage; this is the subject of the next two sections.

3. VARIATION IN THE STRENGTH OF SELECTED CODON USAGE BIAS AMONG BACTERIA

We have previously examined the strength of selected codon usage bias in 80 distinct bacterial species with genome sequences available (Sharp *et al.* 2005). To quantify the strength of selected codon usage bias, we modified a population genetic model (Bulmer 1991). The strength of past selection on codon usage can be estimated from the frequency of optimal codons in a gene, if the expected frequency of those codons in the absence of selection is known. Since, for some amino acids, the identity of the optimal codon varies among species, we focused on four amino acids where it is expected that the same codon would always be favoured by selection. For example, the only Phe tRNA genes known across bacteria have GAA at the anticodon site, and so UUC is always expected to be favoured over UUU, when selection is effective. Similarly, for Tyr, Asn and Ile, G at the critical position of the anticodon should always lead to preference for the C-ending rather than the U-ending codon. To determine the frequency of optimal codons in genes potentially under strong selection, we examined a standard set of 40 highly expressed genes (encoding translation elongation factors and ribosomal proteins) found in all bacterial species; these genes encode proteins with around 10^4–10^5 copies in the *E. coli* cell (Ishihama *et al.* 2008). To define an analogous set of genes present, and expressed at low levels, in all bacteria is more difficult. So we used the codon usage of the genome as a whole as an estimate of the pattern of codon usage when selection is weak; this can be justified because only a minority of genes within a genome are highly expressed. Comparison of codon usage between these two datasets yields an estimate of the compound parameter $S = 2N_e s$, where N_e is the effective population size and s is the selective difference between optimal and non-optimal codons. Thus, S might vary among species because there have been differences in either their population sizes or the strength of selection.

Application of this approach to 80 bacterial genomes revealed considerable variation among species (Sharp *et al.* 2005). The S value for *E. coli* was 1.49. In 24 species (30% of the total), including *H. pylori*, the S value was not significantly greater than zero, providing no evidence for selected codon usage bias. Thirty species (37.5%) had S values greater than 1, with the highest value (2.65) seen in *Clostridium perfringens*, a widespread bacterium that causes a variety of diseases but is most famous as a 'flesh-eating bug'. The 80 species examined included variable numbers of representatives from 14 different major lineages (phyla) of bacteria. There was clear phylogenetic clustering of species with high or low S values, but species with strongly selected codon usage bias occurred in several different phyla. Of the 20 species from the gamma proteobacteria (which includes *E. coli*), nine had S values greater than 1. These nine species form a clade together with a lineage comprised of four species with low S values (figure 1). Thus, it appears that strongly selected codon usage bias evolved on the branch leading to this clade (which includes the orders Enterobacteriales, Pasteurellales, Vibrionales and Alteromonadales) and was subsequently lost on the lineage including *Buchnera* species and *Wigglesworthia*. *Buchnera* species and *Wigglesworthia* are endosymbionts of insects, which have undergone genome reduction and apparently a general relaxation of genomic selection pressures owing to reduced effective population sizes (Moran & Wernegreen 2000; Wernegreen & Funk 2004); one symptom of this is their long branch lengths in the evolutionary tree (figure 1), reflecting an increased rate of molecular evolution.

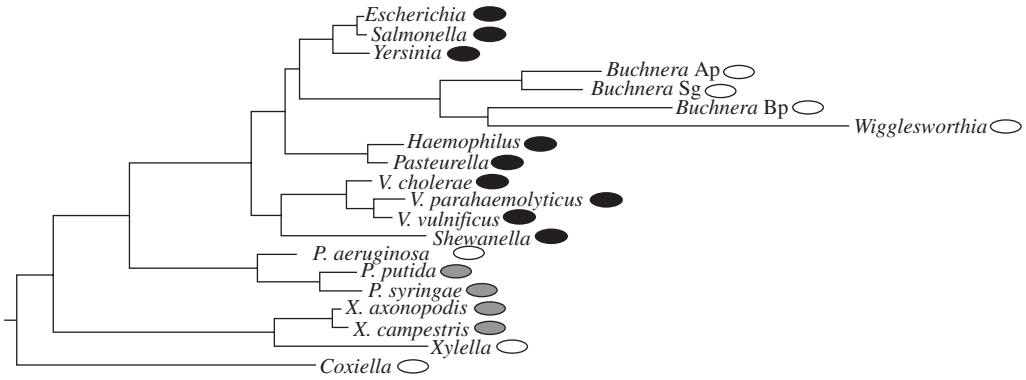


Figure 1. Variation in the strength of selected codon usage bias (S) in gamma proteobacteria. Species are denoted by their genus names, except where there are multiple species from a genus; the abbreviated genus names are *Vibrio*, *Pseudomonas* and *Xanthomonas*. The three *Buchnera* strains are species infecting different aphid hosts. Shaded ovals next to species names indicate the magnitude of S: white ($S < 0.2$), grey ($0.2 < S < 1.0$), black ($S > 1.0$). Phylogenetic relationships and S values were taken from Sharp *et al.* (2005). Note that the clustering of *Wigglesworthia* with *Buchnera* species may be a phylogenetic artefact (Herbeck *et al.* 2005); if so, reduced S values evolved independently in the two lineages.

Across the 80 species, values of S were found to be strongly positively correlated with both the number of rRNA operons and the number of tRNA genes in the genome, even after correction for the underlying phylogenetic relationships among species. Many of the species with low S values had only one rRNA operon and a minimal complement of (around 30–40) tRNA genes. In contrast, the *E. coli* genome has seven rRNA operons and 86 tRNA genes. These results were interpreted as reflecting selection for a co-adapted suite of genomic characteristics required for rapid growth (Sharp *et al.* 2005). For example, *C. perfringens* has 10 rRNA operons and 96 tRNA genes and can replicate in only 7 min under ideal conditions (Labbe & Huang 1995).

To test this association with growth rate, we have used minimum generation time data for 76 of these 80 species, drawn from the compilations made by E.P.C. Rocha (Rocha 2004; Coutourier & Rocha 2006). rRNA operon number, tRNA gene number and S values are all strongly negatively correlated with generation time (figure 2). Using independent contrasts to overcome the fact that the data points are linked by an underlying phylogeny (Felsenstein 1985), the correlation coefficients for rRNA, tRNA and S are 0.35, 0.27 and 0.49, respectively, and all are highly significant ($p < 0.01$). Thus, selection for rapid growth appears to have selected for an increase in the number of rRNA operons and tRNA genes, and for codon usage more strongly biased towards translationally optimal codons.

The observation that closely related species tend to have similar S values (as in figure 1) may reflect similarity of lifestyles, such that closely related bacteria are subject to similar strengths of selection for rapid growth. However, it is also likely that codon usage patterns change relatively slowly. Some of the outlier species in figure 2 could be on lineages that have recently entered a new niche. If a species changed from a lifestyle where rapid growth was advantageous, to one where it was not, it would take some time for strongly selected codon usage bias to decay. That is, the values of S reflect selection on codon usage over a long evolutionary period, but not necessarily the current strength of selection.

4. VARIATION IN THE STRENGTH OF SELECTION ON CODON USAGE BIAS AMONG BACTERIA

An alternative approach, which aims to estimate the strength of *current* selection on codon usage bias, is to examine the frequency spectrum of optimal codons across polymorphic sites. In an equilibrium population, assuming an infinite sites model and free recombination among sites, the effect of selection on the frequency spectrum can be predicted (McVean & Charlesworth 1999). In the absence of selection, the distribution is expected to be U-shaped with a mean of 0.5, but as the strength of selection is increased, the distribution becomes skewed towards higher frequencies of optimal codons. Importantly, this distribution is not expected to be influenced by mutation biases (McVean & Charlesworth 1999). The observed distribution of allele frequencies can be compared to those predicted for different values of $2N_e s$, to obtain the maximum likelihood estimate of this compound parameter, termed gamma (Cutter & Charlesworth 2006). Note that both gamma and S (from the previous section) are estimates of $2N_e s$, but gamma differs from S in reflecting current, ongoing selection. Cutter & Charlesworth (2006) applied this approach to gene sequences from a eukaryote (*Caenorhabditis remanei*), and found a strong correlation between estimates of gamma and the strength of codon usage bias reflecting long-term evolution (as summarized here by S). Here, we use a similar approach to analyse bacterial codon usage.

We applied the method to 25 genome sequences of *E. coli* (including strains of *Shigella* 'species', which lie within the radiation of *E. coli*). First, we analysed polymorphic codon sites in the same 40 highly expressed genes used to estimate S above. All sites with non-synonymous variation, or more than two alleles, were excluded from the analysis, as were sites where the two alleles were both optimal or both non-optimal codons; the latter included sites encoding Cys and

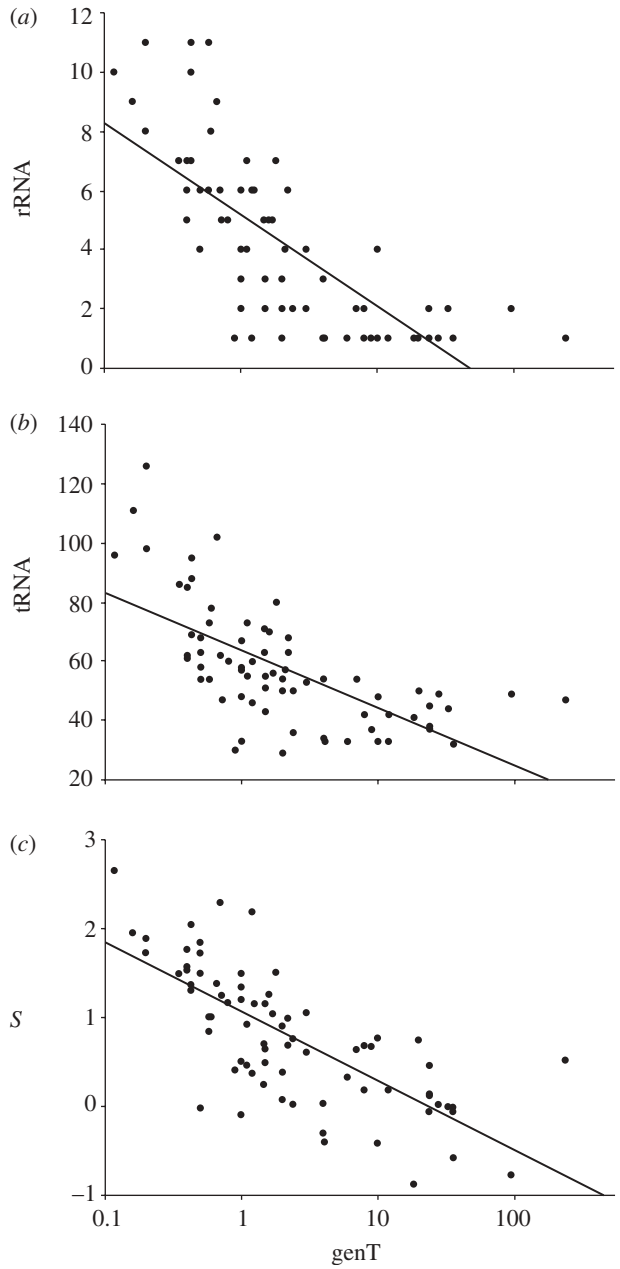


Figure 2. Correlations of (*a*) rRNA operon copy number, (*b*) tRNA gene copy number and (*c*) the strength of selected codon usage bias (*S*), with generation time in bacteria. The minimum generation time (in hours) is plotted on a logarithmic scale.

Lys, where no optimal codon was designated (table 1). Among nearly 6000 potentially synonymously variable codons, 194 were segregating for one optimal and one non-optimal codon. The average frequency of optimal codons across the polymorphic sites (q_{opt}) was well in excess of 0.5 and the gamma value was estimated as 1.70 (table 2); this value is not substantially (and certainly not significantly) different from the *S* value of 1.49 estimated by comparing the codon usage of these 40 genes to that in the genome as a whole.

We then examined 10 genes with low codon usage bias and expressed at low levels. From the 20 chromosomal genes encoding proteins with the lowest recorded copy numbers (between 50 and 100 per cell) in Ishihama *et al.* (2008), we selected those having codon adaptation index (CAI) values between 0.3 and 0.4. The CAI (Sharp & Li 1987) is a widely used species-specific measure of selected codon usage bias, which would take a maximum value of 1 for a gene using only optimal codons. In *E. coli* K-12, the range (99th percentile) of CAI values is from 0.15 to 0.74, with a median of 0.31. Thus, the 10 genes selected here do not have the lowest CAI values, but a substantial fraction of genes with lower values are either hypothetical or of likely foreign origin (i.e. owing to horizontal gene transfer). These 10 'low' genes exhibited a much higher level of polymorphism for optimal versus non-optimal codons, consistent with much lower levels of constraint on codon usage in these genes; genes with lower codon usage bias also exhibit higher levels of interspecific divergence at synonymous sites (Sharp *et al.* 1989; but see also Berg & Martelius 1995; Eyre-Walker & Bulmer 1995). The average frequency of optimal codons among polymorphic sites across the 25 strains was very close to the value of 0.5 expected in the absence of selection and consequently the estimated gamma value was very close to zero (table 2).

In contrast, we analysed seven genome sequences of *H. pylori*, where codon selection has previously been estimated to be very weak ($S = 0.02$; Sharp *et al.* 2005). We focused on the same 40 highly expressed genes as above, where selection (if present) should be strongest. In this species there is a difficulty identifying which codons are optimal, because there is little difference between the codon usage of highly expressed and other genes (Lafay *et al.* 2000). Therefore, we focused on the four amino acids used to derive *S* values, where the C-ending codons are expected to be optimal for biochemical reasons, even if they are not preferred because selection is ineffective. In contrast to the analysis of 40 highly expressed genes in *E. coli*, the average frequency of optimal codons across polymorphic sites was only just greater than 0.5, and the estimate of gamma was not significantly greater than zero (table 2).

Finally, we examined the sequences of five genes determined for 247 strains of *C. perfringens* (Rooney *et al.* 2006). Sixteen optimal codons for *C. perfringens* were defined by the same approach as applied to *E. coli* in table 1. The five genes vary in their strength of codon usage bias (measured by F_{op} in table 3), apparently reflecting differing levels of expression. These data are limited in terms of the number of polymorphic sites. Nevertheless, values of F_{op} and of the average frequency of optimal codons across polymorphic sites showed the same rank order across the five genes (table 3). Similar to *E. coli*, polymorphic sites in genes with low codon usage bias had average frequencies of optimal codons close to 0.5, yielding gamma values close to zero. The most highly expressed gene in the dataset, *rplL*, is one of the 40 genes in the highly expressed datasets used above, and seems representative of that dataset because its F_{op} value is very close to that obtained from the 40 genes as a whole ($F_{op} = 0.647$). The estimated gamma value for *rplL* was 3.28; the value has very wide confidence intervals reflecting the small number of polymorphic sites, but again it is quite close to the *S* value of 2.65 estimated for this species.

Table 2. Estimates of the strength of selection for optimal codons from polymorphism data from *E. coli* and *H. pylori*.

species[a]	S[b]	genes[c]	CAI[d]	sites[e]	poly[f]	q_{opt}[g]	gamma[h] (95% CI)
E. coli ($n = 25$)	1.49	40 high	0.67	5963	194	0.69	1.70 (1.23 to 2.25)
		10 low	0.35	7211	1255	0.50	−0.04 (−0.20 to 0.12)
H. pylori ($n = 7$)	0.02	40 high	n.a.	1131	172	0.52	0.28 (−0.27 to 0.84)

[a]All complete genome sequences were obtained from the GenBank database (July 2009). For *E. coli*, where multiple substrains of one strain were present, only one was retained.
[b]S is an estimate of the strength of selection on optimal codons ($=2N_e s$) from comparisons of codon usage bias.
[c]The 40 high genes are a standard set of 40 genes expressed at high levels in all species (Sharp *et al.* 2005). The 10 low genes are 10 genes expressed at low levels in *E. coli* (*dld*, *glnE*, *helD*, *metL*, *mutS*, *rmuC*, *spoT*, *uvrD*, *yebT*, *yhdP*).
[d]The CAI is a species-specific measure of selected codon usage bias (Sharp & Li 1987), here calculated for a concatenation of the genes.
[e]The number of potentially synonymously variable sites examined.
[f]The number of sites polymorphic for optimal and non-optimal codons.
[g]The average frequency of optimal codons at sites polymorphic for optimal and non-optimal codons.
[h]Gamma is an estimate (with 95% CI) of the strength of selection on optimal codons ($=2N_e s$) from polymorphism data.

Table 3. Estimates of the strength of selection (gamma) for optimal codons from polymorphism among 247 strains of *C. perfringens*.

gene	F_{op}[a]	sites[b]	q_{opt}[c]	gamma[d] (95% CI)
rplL	0.65	6	0.88	3.28 (0.48 to 8.77)
gyrA	0.49	25	0.76	1.74 (0.62 to 3.06)
colA	0.42	33	0.58	0.48 (−0.37 to 1.37)
plc	0.39	47	0.56	0.34 (−0.37 to 1.07)
pfoS	0.35	22	0.48	−0.10 (−1.15 to 0.95)

[a]The frequency of *C. perfringens* optimal codons in one copy of the gene sequence.
[b]The number of sites polymorphic for optimal and non-optimal codons.
[c]The average frequency of optimal codons at sites polymorphic for optimal and non-optimal codons.
[d]Gamma is an estimate (with 95% CI) of the strength of selection on optimal codons ($=2N_e s$) from polymorphism data.

These analyses of the frequency spectrum of optimal codons across polymorphic sites should be taken with caution, since the approach assumes that the sequences are drawn randomly from an interbreeding population at mutation-selection-drift equilibrium (McVean & Charlesworth 1999). It has been shown that a recent change in population size can have an erratic impact on the expected frequencies (Zeng & Charlesworth 2009), but the apparent consistency between the values of $2N_e s$ estimated by gamma and by S suggests that such issues have not been important in the examples considered here. Furthermore, the frequency spectra for the two *E. coli* datasets analysed here, with 25 sequences and estimated gamma values of 0 and 1.7 (figure 3), appear (qualitatively) remarkably similar to the expected distributions for samples of 20 sequences from a diploid species, with gamma values of 0 and 4, shown by McVean & Charlesworth (1999).

The effect of population subdivision on the frequency spectrum, which may be particularly relevant to bacterial species, has not been investigated in detail. In a goodness-of-fit test, the site frequency spectrum for the *E. coli* low-expression genes differed significantly from that expected ($\chi^2 = 50.1$, d.f. $= 22$, $p < 10^{-3}$), largely because of the excess of sites with an optimal allele frequency of 7 (figure 3). Such an excess of sites with optimal codons segregating at intermediate frequencies might be expected in samples drawn from a subdivided population. However, the effect is quite small, suggesting again that, for these data, the extent to which the real populations violate the assumptions of the model has had little impact on the results.

The main discrepancy between the observed and expected distributions concerns an excess of sites at extreme optimal codon frequencies in the highly expressed genes; i.e. the leftmost and rightmost grey columns in figure 3 are taller than would be predicted. A possible explanation for the excess of sites with low optimal codon frequencies is that there are certain sites where a codon that is normally optimal is not advantageous. This may be related to the context of the codon: while overall, the frequency of GAA rather than GAG for Glu is only increased a little in highly expressed genes in *E. coli* (table 1), it has been found that the preference for GAA is strong when the following codon begins with G, but weak in other contexts (Maynard Smith & Smith 1996; Berg & Silva 1997). Also, near the start of highly expressed genes in *E. coli*, the use of optimal codons is reduced and the frequency of A-ending codons is unusually high, seemingly reflecting conflicting selection pressures (Eyre-Walker & Bulmer 1993). Although we saw no obvious peculiarities to the sites where non-optimal codons were segregating at high frequencies, this merits further investigation.

5. DISCUSSION

The extensive variation in codon usage patterns seen among bacteria is most likely primarily owing to differences in mutation biases. However, in many—but not all—species there is additional variation among genes that is consistent with the action of natural selection. The observation that genes expressed at high levels have increased frequencies of those codons that are expected to be translationally optimal is strongly suggestive that these codons are selectively favoured. The fact that, for some amino acids, the identity of the optimal codons differs among species, coordinated with changes in the population of tRNA genes, reinforces the view that this bias in codon usage is adaptive. However, numerous aspects of how and

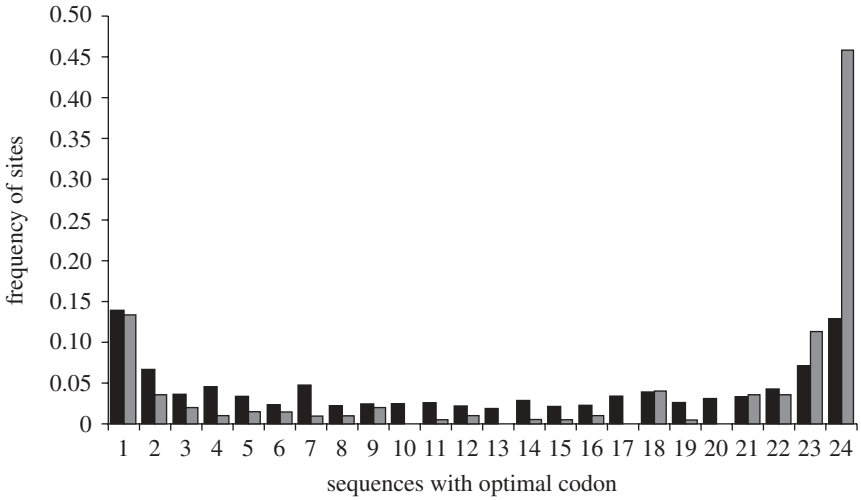


Figure 3. Distribution of the number of optimal codons at polymorphic sites in 25 strains of *E. coli*. Data are presented for two sets of genes: 10 genes expressed at low levels (black) and 40 genes expressed at high levels (grey; see also table 2).

why natural selection has shaped patterns of codon usage remain unresolved.

To learn more about this we have applied two different approaches to estimate the strength of selection on codon usage in bacteria. Both methods provide estimates of $2N_es$, which compounds the effective population size with the selective difference between optimal and non-optimal codons. However, the values (S) from one method reflect very long-term evolution, whereas those (gamma) from the other reflect ongoing selection at polymorphic sites. Gamma values might be expected to be particularly sensitive to the assumption that the sequences analysed came from an idealized population, but the gamma values estimated here were remarkably consistent with estimates of S from the same species. This should not always be the case. Across the phylogeny of bacteria, there have probably been many instances when selection pressures have changed. Then, for example, in a species where codon selection has recently stopped, the gamma value may be close to zero but the S value may still be high because biased codon usage may take a long time to decay (Lawrence & Ochman 1997). In addition, recent demographic changes may impact on the frequency of optimal codons at polymorphic sites and hence gamma (Zeng & Charlesworth 2009), without significant impact on S. At the extreme, *Yersinia pestis* (the cause of plague) appears to have gone through a very recent severe bottleneck (Achtman *et al.* 2004) and has such little nucleotide diversity that it would be very difficult to estimate gamma; however, its S value is 1.15 (Sharp *et al.* 2005), predominantly reflecting selection that occurred in the past in the species from which it was derived, *Y. pseudotuberculosis*.

Highly expressed genes in many bacteria have S values around 1. The magnitude of the selective difference between optimal and non-optimal codons would then be estimated as on the order of the reciprocal of the effective population size. The effective population sizes for bacterial species are probably not known with any accuracy, but typical values might be in the order of 10^8 (Lynch 2007), implying that the fitness difference between an optimal and a non-optimal codon may be around 10^{-8}. This is a miniscule value, perhaps reflecting the most subtle form of natural selection known, and only estimable because the selection is repeated over many sites.

This tiny selection coefficient raises a number of issues. One is whether the same form of codon selection can be operating in multicellular eukaryotes. The same approach to estimate S has been adapted for application in eukaryotes (dos Reis & Wernisch 2009). For *D. melanogaster* and *Caenorhabditis elegans*, S values of 1.08 and 1.96 were obtained. The same approach to estimate gamma values has been applied to *Caenorhabiditis remanei* yielding an average value of 0.44 across genes, but with values greater than 1.0 in some genes (Cutter 2008). A number of other analyses have used alternative methods to estimate N_es for codon usage from polymorphism data of *Drosophila* species. These methods usually require an assignment of the ancestral state at a polymorphic site, which may be difficult in some cases, and especially error prone with bacteria; hence we did not use them here. These analyses have also yielded estimates of the same order of magnitude; for example, Maside *et al.* (2004) estimated N_es to be around 0.65 in *D. americana*. Thus, estimates of N_es for *Drosophila* and *Caenorhabditis* are of the same order of magnitude as those for bacteria. However, estimates of N_e are typically two orders of magnitude lower than the value given above for bacteria, implying that the fitness difference associated with optimal codons must be two orders of magnitude larger. This has led Lynch (2007) to question whether codon bias in these eukaryotes is caused by some other force, such as biased gene conversion, rather than selection.

A second issue concerns the many sites in the genome where selection on codon usage has occurred. Linkage between sites impairs the efficacy of selection on any one of them, analogous to reducing the effective population size (Hill & Robertson 1966). Bacteria typically have one relatively small chromosome, in which all of the highly expressed genes are linked, and so the strength of selected codon usage bias is expected to be reduced (Li 1987; McVean &

Charlesworth 2000). Nevertheless, bacteria have various means of recombination, which vary in frequency among species. This variation in recombination rates could influence the strength of selected codon usage bias, although it has apparently not impacted on *H. pylori*, which has $N_e s$ close to zero, despite perhaps the highest rate of recombination known among bacteria (Suerbaum *et al*. 1998). If the various sites under codon selection in *Drosophila* and *Caenorhabditis* are much less tightly linked than those in bacteria, this could contribute to easing the paradox of similar estimates of $N_e s$ in eukaryotes and bacteria (Kaiser & Charlesworth 2009).

The reason why translationally optimal codons are advantageous is also unresolved: it is assumed that they can enhance translational efficiency and/or translational accuracy, but which is more important? The observation that variation among bacteria in the strength of selected codon usage bias is strongly correlated with growth rate (figure 2) may bear on at least one aspect of this debate. In arguing for the accuracy hypothesis, Drummond & Wilke (2008) suggested that non-optimal codons decrease fitness because mistranslated proteins can be toxic. Under this hypothesis, selection against non-optimal codons is stronger in more highly expressed genes because they have more opportunity to be mistranslated. However, it is not clear that the toxic effect of mistranslated proteins would be dependent on rapid growth rate. In contrast, it is obvious that the observed correlation of $N_e s$ with growth rate is consistent with the efficiency hypothesis. However, this does not rule out the possibility that inaccuracy of translation is selected against because of its negative impact on the efficiency of translation (Bulmer 1991).

Finally, given the observation that (in many species) codon usage in highly expressed genes is strongly selected and matches tRNA abundance, and yet the identity of the optimal codons can vary among species, there remains an intriguing question: how can this state of co-adaptation between the tRNA gene complement and the codon usage bias in highly expressed genes diverge over time? The observation that the strength of selection varies greatly among contemporary species suggests that there could have been times when ancestral species were subject to relaxed selection, due perhaps to a change of lifestyle or greatly reduced effective population size; selected codon usage bias would then drift and decay. After the re-imposition of selection pressure, the genome could then move to a co-adapted state different from that in the original ancestor. Alternatively, Shields (1990) has suggested that a prolonged influence of mutation bias could provide the impetus for a shift, without the need for a period of drift. Detailed analyses of switches in the identity of optimal codons across the phylogeny of bacteria may provide insights into which, if either, of these processes has played a major role in shaping the patterns of selected codon usage seen in bacteria.

We are indebted to Brian Charlesworth for discussion of various aspects of this topic. This work was supported by a studentship from the UK Biotechnology and Biological Sciences Research Council to L.R.E. and a Royal Society of Edinburgh/Caledonian Research Foundation Biomedical Personal Research Fellowship to K.Z.

REFERENCES

Achtman, M. *et al.* 2004 Microevolution and history of the pague bacillus, Yersinia pestis. *Proc. Natl Acad. Sci. USA* **101**, 17 837–17 842. (doi:10.1073/pnas.0408026101)

Akashi, H. 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935.

Andersson, S. G. E. & Kurland, C. G. 1990 Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54**, 198–210.

Belozersky, A. N. & Spirin, A. S. 1958 A correlation between the compositions of deoxyribonucleic and ribonucleic acids. *Nature* **182**, 111–112. (doi:10.1038/182111a0)

Berg, O. G. & Martelius, M. 1995 Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. *J. Mol. Evol.* **41**, 449–456. (doi:10.1007/BF00160316)

Berg, O. G. & Silva, P. J. N. 1997 Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. *Nucleic Acids Res.* **25**, 1397–1404. (doi:10.1093/nar/25.7.1397)

Bernardi, G. & Bernardi, G. 1986 Compositional constraints and genome evolution. *J. Mol. Evol.* **24**, 1–11. (doi:10.1007/BF02099946)

Bibb, M. J., Findlay, P. R. & Johnson, M. W. 1984 The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* **30**, 157–166. (doi:10.1016/0378-1119(84)90116-1)

Brewer, B. J. 1988 When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**, 679–686. (doi:10.1016/0092-8674(88)90086-4)

Bulmer, M. 1991 The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L. & McAdams, H. H. 2004 Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl Acad. Sci. USA* **101**, 3480–3485. (doi:10.1073/pnas.0307827100)

Coutourier, E. & Rocha, E. P. C. 2006 Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* **59**, 1506–1518. (doi:10.1111/j.1365-2958.2006.05046.x)

Cutter, A. D. 2008 Multilocus patterns of polymorphism and selection across the X chromosome of *Caenorhabditis remanei*. *Genetics* **178**, 1661–1672. (doi:10.1534/genetics.107.085803)

Cutter, A. D. & Charlesworth, B. 2006 Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Curr. Biol.* **16**, 2053–2057. (doi:10.1016/j.cub.2006.08.067)

dos Reis, M. & Wernisch, L. 2009 Estimating translational selection in eukaryotic genomes. *Mol. Biol. Evol.* **26**, 451–461. (doi:10.1093/molbev/msn272)

Drummond, D. A. & Wilke, C. O. 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352. (doi:10.1016/j.cell.2008.05.042)

Duret, L. & Mouchiroud, D. 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila* and *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **96**, 4482–4487. (doi:10.1073/pnas.96.8.4482)

Ehrenberg, M. & Kurland, C. G. 1984 Costs of accuracy determined by a maximal growth rate constraint. *Quart. Rev. Biophys.* **17**, 45–82. (doi:10.1017/S0033583500005254)

Eyre-Walker, A. & Bulmer, M. 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* **21**, 4599–4603. (doi:10.1093/nar/21.19.4599)

Eyre-Walker, A. & Bulmer, M. 1995 Synonymous substitution rates in Enterobacteria. *Genetics* **140**, 1407–1412.

Felsenstein, J. 1985 Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15. (doi:10.1086/284325)

Fijalkowska, I. J., Jonczyk, P., Tkaczyk, M. M., Bialoskorska, M. & Schaaper, R. M. 1998 Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proc. Natl Acad. Sci. USA* **95**, 10 020–10 025. (doi:10.1073/pnas.95.17.10020)

Francino, M. P., Chao, L., Riley, M. A. & Ochman, H. 1996 Asymmetries generated by transcription-coupled repair in Enterobacterial genes. *Science* **272**, 107–109. (doi:10.1126/science.272.5258.107)

Galtier, N. & Lobry, J. R. 1997 Relationships between genomic G + C content, RNA secondary structures, and optimal growth temperatures in prokaryotes. *J. Mol. Evol.* **44**, 632–636. (doi:10.1007/PL00006186)

Gouy, M. & Gautier, C. 1982 Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**, 7055–7074. (doi:10.1093/nar/10.22.7055)

Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. 1980 Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**, r49–r62.

Henry, I. & Sharp, P. M. 2007 Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.* **24**, 10–12. (doi:10.1093/molbev/msl148)

Herbeck, J. T., Degnan, P. H. & Wernegreen, J. J. 2005 - Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the Enterobacteriales (gamma-Proetobacteria). *Mol. Biol. Evol.* **22**, 520–532. (doi:10.1093/molbev/msi036)

Hill, W. G. & Robertson, A. 1966 The effect of linkage in limits to artificial selection. *Genet. Res.* **8**, 269–294. (doi:10.1017/S0016672300010156)

Ikemura, T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M. J. & Frishman, D. 2008 Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genom.* **9**, 102. (doi:10.1186/1471-2164-9-102)

Kaiser, V. B. & Charlesworth, B. 2009 The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.* **25**, 9–12. (doi:10.1016/j.tig.2008.10.009)

Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T. 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 143–155. (doi:10.1016/S0378-1119(99)00225-5)

Kloster, M. & Tang, C. 2008 SCUMBLE: a method for systematic and accurate detection of codon usage bias by maximum likelihood estimation. *Nucleic Acids Res.* **36**, 3819–3827. (doi:10.1093/nar/gkn288)

Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. 2009 Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258. (doi:10.1126/science.1170160)

Labbe, R. G. & Huang, T. H. 1995 Generation times and modeling of enterotoxin-positive and enterotoxin-negative strains of *Clostridium perfringens* in laboratory media and ground beef. *J. Food Prot.* **58**, 1303–1306.

Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M. & Wolfe, K. H. 1999 Proteome composition and codon usage in spirochaetes, species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* **27**, 1642–1649. (doi:10.1093/nar/27.7.1642)

Lafay, B., Atherton, J. C. & Sharp, P. M. 2000 Absence of translationally selected codon usage bias in *Helicobacter pylori*. *Microbiology* **146**, 851–860.

Lawrence, J. G. & Ochman, H. 1997 Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**, 383–397. (doi:10.1007/PL00006158)

Li, H. 1987 Models of nearly neutral mutations with particular implications for the nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345. (doi:10.1007/BF02134132)

Lobry, J. R. 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665.

Lynch, M. 2007 *The origins of genome architecture*. MA, USA: Sinauer Associates.

Lynn, D. J., Singer, G. A. C. & Hickey, D. A. 2002 Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30**, 4272–4277. (doi:10.1093/nar/gkf546)

Maside, X., Lee, A. W. & Charlesworth, B. 2004 Selection on codon usage in *Drosophila americana*. *Curr. Biol.* **14**, 150–154. (doi:10.1016/j.cub.2003.12.055)

Maynard Smith, J. & Smith, N. H. 1996 Site-specific codon bias in bacteria. *Genetics* **142**, 1037–1043.

McLean, M., Wolfe, K. H. & Devine, K. M. 1998 Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* **47**, 691–696. (doi:10.1007/PL00006428)

McVean, G. A. T. & Charlesworth, B. 1999 A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* **74**, 145–158. (doi:10.1017/S0016672399003912)

McVean, G. A. T. & Charlesworth, B. 2000 The effects of Hill–Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**, 929–944.

Moran, N. A. & Wernegreen, J. J. 2000 Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol. Evol.* **15**, 321–326. (doi:10.1016/S0169-5347(00)01902-9)

Musto, H., Romero, H. & Zavala, A. 2003 Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*. *Microbiology* **149**, 855–863.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. & Bernardi, G. 2004 Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* **573**, 73–77. (doi:10.1016/j.febslet.2004.07.056)

Muto, A. & Osawa, S. 1987 The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA* **84**, 166–169. (doi:10.1073/pnas.84.1.166)

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A. & Hattori, M. 2006 The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**, 267. (doi:10.1126/science.1134196)

Normand, P. *et al.* 2007 Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Res.* **17**, 7–15. (doi:10.1101/gr.5798407)

Pedersen, S., Bloch, P. L., Reeh, S. & Neidhardt, F. C. 1978 Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. *Cell* **14**, 179–190. (doi:10.1016/0092-8674(78)90312-4)

Post, L. E. & Nomura, M. 1980 DNA sequences from the *str* operon of *Escherichia coli*. *J. Biol. Chem.* **255**, 4660–4666.

Rocha, E. P. C. 2004 Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient coding for translation optimization. *Genome Res.* **14**, 2279–2286. (doi:10.1101/gr.2896904)

Rooney, A. P., Swezey, J. L., Friedman, R., Hecht, D. W. & Maddox, C. W. 2006 Analysis of core housekeeping and virulence genes reveals cryptic lineages of *Clostridium perfringens* that are associated with distinct disease presentations. *Genetics* **172**, 2081–2092. (doi:10.1534/genetics.105.054601)

Sharp, P. M. 1991 Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*, codon usage, map position and concerted evolution. *J. Mol. Evol.* **33**, 23–33. (doi:10.1007/BF02100192)

Sharp, P. M. & Li, W.-H. 1987 The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295. (doi:10.1093/nar/15.3.1281)

Sharp, P. M., Shields, D. C., Wolfe, K. H. & Li, W.-H. 1989 Chromosomal location and evolutionary rate variation in Enterobacterial genes. *Science* **246**, 808–810. (doi:10.1126/science.2683084)

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. 2005 Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**, 1141–1153. (doi:10.1093/nar/gki242)

Shields, D. C. 1990 Switches in species-specific codon preferences: the influence of mutation biases. *J. Mol. Evol.* **31**, 71–80. (doi:10.1007/BF02109476)

Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. 1988 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among alternative synonymous codons. *Mol. Biol. Evol.* **5**, 704–716.

Sørensen, M. A. & Pedersen, S. 1991 Absolute in vivo translation rates of individual codons in *Escherichia coli*: the two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.* **222**, 265–280. (doi:10.1016/0022-2836(91)90211-N)

Stoletzki, N. & Eyre-Walker, A. 2006 Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* **24**, 374–381. (doi:10.1093/molbev/msl166)

Suerbaum, S., Maynard Smith, J., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. 1998 Free recombination within *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **95**, 12 619–12 624. (doi:10.1073/pnas.95.21.12619)

Sueoka, N. 1962 On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592.

Tillier, E. R. M. & Collins, R. A. 2000 The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50**, 249–257.

Wernegreen, J. J. & Funk, D. J. 2004 Mutation exposed: a neutral explanation for extreme base composition of an endosymbiont genome. *J. Mol. Evol.* **59**, 849–858. (doi:10.1007/s00239-003-0192-z)

Zeng, K. & Charlesworth, B. 2009 Estimating selection intensity on synonymous codon usage in a non-equilibrium population. *Genetics* **183**, 651–662. (doi:10.1534/genetics.109.101782)

biology letters
Molecular evolution

Biol. Lett.
doi:10.1098/rsbl.2010.0620
Published online

Impact of translational selection on codon usage bias in the archaeon *Methanococcus maripaludis*

Laura R. Emery and Paul M. Sharp*

Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK
**Author for correspondence* (*paul.sharp@ed.ac.uk*).

Patterns of codon usage have been extensively studied among Bacteria and Eukaryotes, but there has been little investigation of species from the third domain of life, the Archaea. Here, we examine the nature of codon usage bias in a methanogenic archaeon, *Methanococcus maripaludis*. Genome-wide patterns of codon usage are dominated by a strong A + T bias, presumably largely reflecting mutation patterns. Nevertheless, there is variation among genes in the use of a subset of putatively translationally optimal codons, which is strongly correlated with gene expression level. In comparison with Bacteria such as *Escherichia coli*, the strength of selected codon usage bias in highly expressed genes in *M. maripaludis* seems surprisingly high given its moderate growth rate. However, the pattern of selected codon usage differs between *M. maripaludis* and *E. coli*: in the archaeon, strongly selected codon usage bias is largely restricted to twofold degenerate amino acids (AAs). Weaker bias among the codons for fourfold degenerate AAs is consistent with the small number of tRNA genes in the *M. maripaludis* genome.

Keywords: codon usage; translation; selection; Archaea

1. INTRODUCTION

The frequencies of alternative synonymous codons vary among species and among genes within a genome, reflecting the combined effects of mutation bias, natural selection and random genetic drift [1,2]. Selection primarily favours translationally optimal codons—those best recognized by the most abundant tRNA species [3]. Selection has the greatest impact on highly expressed (HE) genes, because their translation has the largest effect on cellular efficiency during competitive growth [4]. The strength of selected codon usage bias varies among species, and in some it is weak or absent. Across Bacteria, the strength of selected bias is positively correlated with the copy numbers of rRNA and tRNA genes, and negatively correlated with generation time [5–8], implying a co-adapted suite of genome characteristics necessary for achieving fast growth rates.

While codon usage has been extensively studied in Bacteria, as well as in some groups of Eukaryotes, there has been little work on patterns of codon bias in the third domain of life, the Archaea. Karlin *et al.* [9] used codon-usage analyses to predict HE genes in 19 archaeal genomes. However, their approach (i) assumed translational selection without first testing for its presence and (ii) is expected to give anomalous results, even in species where selection has shaped codon usage [10]. Some Archaea have been included in larger analyses mainly focused on genomes from Bacteria. For example, Dos Reis *et al.* [11] claimed that six of 16 species of Archaea showed evidence of translational selection, but (at least among Bacteria) their measure of selection does not correlate well with a population genetics-based approach [6].

Here, we investigate variation in patterns of synonymous codon usage across the genes of *Methanococcus maripaludis*, one of the most extensively studied archaeal species. *Methanococcus maripaludis* is a mesophilic methanogen isolated from salt marsh sediment [12]. A complete genome sequence [13] and genome-wide expression data [14] have been determined for this species. *Methanococcus maripaludis* has a fastest doubling time of 2.3 h at 37°C [12], a near-minimal complement of tRNA genes, and only three rRNA operons. These features seem typical of many Archaea and, by comparison with Bacteria, it might be predicted that such species should have little or no selected codon usage bias. Nevertheless, we find clear evidence of translational selection in HE genes in *M. maripaludis*, although the pattern of codon bias is rather different from those seen in most Bacteria.

2. MATERIAL AND METHODS

Protein-coding sequences from the genome of *M. maripaludis* strain S2 [13] were obtained from the GenBank database (accession no. BX950229) using the ACNUC retrieval system [15]. Six genes with fewer than 50 codons were excluded from subsequent analyses. The numbers of tRNA genes, with their predicted anti-codon sequences, were obtained from the tRNA scan SE database [16]. Gene expression level was estimated from protein abundance data [14], as the signal intensity (n_1 values) normalized by protein molecular weight.

For each gene the deviation from random codon usage was measured using the effective number of codons, N_c [17]; N_c values potentially range from 20, when only one synonym is used for each amino acid (AA), to 61 when all codons are used randomly. A plot of N_c values against GC3s, the G + C content at synonymously variable third positions of codons, is useful to explore variation in patterns of codon bias. Within-group correspondence analysis was also used to identify any major trends among genes (see electronic supplementary material).

A dataset of 51 expected HE genes encoding ribosomal proteins was identified on the basis of genome annotation, expression level data [14] and conservation across Archaea (see electronic supplementary material). Optimal codons were identified as those occurring significantly more frequently in the HE genes than across all genes, using χ^2-tests with sequential Bonferroni correction [10]. Codon adaptation index (CAI; [18]) values were computed using codon fitness values from the HE gene set. Finally, the strength of selected codon usage bias (S) in the HE genes was estimated with the method used for Bacteria by Sharp *et al.* [6]; the method was adapted to obtain analogous values for individual AAs in both *M. maripaludis* and *Escherichia coli*.

3. RESULTS

(a) *Variation in codon bias among genes in* Methanococcus maripaludis

The A + T richness of the *M. mariplaudis* genome dominates its overall codon usage (table 1), with an

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2010.0620 or via http://rsbl.royalsocietypublishing.org.

Received 7 July 2010
Accepted 12 August 2010

Table 1. Codon usage in *Methanococcus maripaludis*, summed across across all genes (all) and for 51 HE genes (high). The second value in each case is the relative synonymous codon usage (the observed number divided by the average for that AA).

		high		all				high		all				high		all				high		all	
Phe	UUU	72	0.75	16 089	1.56	Ser	UCU	28	0.45	6178	1.21	Tyr	UAU	20	0.21	9940	1.06	Cys	UGU	33	0.85	4047	1.21
Phe	UUC[a]	121	1.25	4522	0.44	Ser	UCC	42	0.67	2430	0.48	Tyr	UAC[a]	171	1.79	8795	0.94	Cys	UGC[a]	45	1.15	2633	0.79
Leu	UUA[a]	304	3.64	18 877	2.53	Ser	UCA[a]	182	2.90	11 554	2.26	Ter	UAA	50	2.94	1460	2.54	Ter	UGA	0	0.00	134	0.23
Leu	UUG	66	0.79	5553	0.74	Ser	UCG	4	0.06	2102	0.41	Ter	UAG	1	0.06	130	0.23	Trp	UGG	44	1.00	3186	1.00
Leu	CUU	58	0.69	12 587	1.69	Pro	CCU[a]	143	1.89	5691	1.37	His	CAU	23	0.34	3105	0.84	Arg	CGU	1	0.01	799	0.32
Leu	CUC	53	0.63	3383	0.45	Pro	CCC	4	0.05	1465	0.35	His	CAC[a]	114	1.66	4255	1.16	Arg	CGC	0	0.00	251	0.10
Leu	CUA	13	0.16	2426	0.33	Pro	CCA	150	1.98	7671	1.84	Gln	CAA[a]	133	1.42	4631	1.01	Arg	CGA	5	0.06	1338	0.53
Leu	CUG	7	0.08	1939	0.26	Pro	CCG	6	0.08	1840	0.44	Gln	CAG	54	0.58	4546	0.99	Arg	CGG	1	0.01	641	0.25
Ile	AUU	225	1.40	23 812	1.55	Thr	ACU	87	1.00	8311	1.36	Asn	AAU	51	0.36	17 956	1.30	Ser	AGU	54	0.86	5047	0.99
Ile	AUC[a]	195	1.21	8136	0.53	Thr	ACC	63	0.72	3585	0.59	Asn	AAC[a]	229	1.64	9706	0.70	Ser	AGC[a]	67	1.07	3383	0.66
Ile	AUA	62	0.39	14 168	0.92	Thr	ACA[a]	189	2.17	9532	1.57	Lys	AAA[a]	803	1.90	38 675	1.78	Arg	AGA[a]	440	5.45	8771	3.47
Met	AUG	207	1.00	12 597	1.00	Thr	ACG	9	0.10	2934	0.48	Lys	AAG	43	0.10	4709	0.22	Arg	AGG	37	0.46	3376	1.33
Val	GUU[a]	361	2.33	17 585	2.05	Ala	GCU[a]	293	1.82	6944	0.98	Asp	GAU	164	1.15	19 381	1.41	Gly	GGU[a]	227	1.50	8138	0.98
Val	GUC	25	0.16	2199	0.26	Ala	GCC	12	0.07	1714	0.24	Asp	GAC[a]	122	0.85	8095	0.59	Gly	GGC	76	0.50	4103	0.50
Val	GUA	208	1.34	12 102	1.41	Ala	GCA	322	2.00	17 607	2.47	Glu	GAA	514	1.87	36 441	1.83	Gly	GGA	285	1.89	17 312	2.09
Val	GUG	26	0.17	2439	0.28	Ala	GCG	16	0.10	2221	0.31	Glu	GAG	36	0.13	3438	0.17	Gly	GGG	16	0.11	3571	0.43

[a]The codons occurring at significantly higher frequencies in the HE dataset.

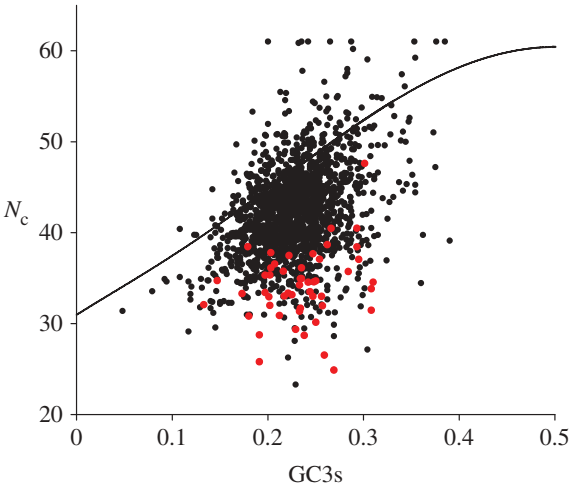


Figure 1. The effective number of codons, N_c, and G + C content at synonymously variable third codon positions, GC3s, for *Methanoccus maripaludis* genes. The line indicates the expected N_c value with random codon usage. The subset of 51 HE genes is shown in red.

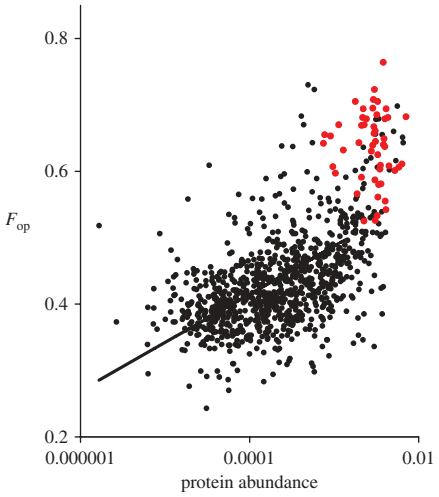


Figure 2. The frequency of optimal codons in a gene (F_{op}) as a function of gene expression level, as estimated from protein abundance data (from [14]). The subset of 51 HE genes is shown in red.

Table 2. The strength of selected codon usage bias (S) in HE genes, for each AA in *M. maripaludis* and *E. coli*. Only AAs with twofold or fourfold degeneracy, and only those for which both species exhibit preference for an optimal codon (shown), are included.

twofold degenerate AAs					fourfold degenerate AAs				
	M. maripaludis		*E. coli*			*M. maripaludis*		*E. coli*	
AA	codon	S	codon	S	AA	codon	S	codon	S
Phe	UUC	1.79	UUC	1.79	Pro	CCU	0.55	CCG	0.79
Tyr	UAC	2.27	UAC	1.39	Thr	ACA	0.61	ACU	1.52
His	CAC	1.28	CAC	1.16	Val	GUU	0.28	GUU	1.14
Asn	AAC	2.12	AAC	1.74	Ala	GCU	0.95	GCU	1.49
Asp	GAC	0.58	GAC	1.17	Gly	GGU	0.61	GGU	1.21

average G + C content at synonymously variable third positions (GC3s) of 0.23. The standard deviation of GC3s values across the 1716 genes analysed (0.040) is only a little higher than the value expected from random binomial variation (0.030). In contrast, N_c values range quite widely (figure 1), indicating variability owing to some additional source(s) of bias. Nearly all of the expected HE genes have very low N_c values given their GC3s, indicating more strongly biased codon usage. Correspondence analysis indicates that there is a single major trend among genes, with the HE genes lying towards one extreme of that trend (see electronic supplementary material).

Eighteen codons occur at significantly higher frequencies in the HE genes than in coding sequences as a whole (table 1). The frequency of these putative optimal codons (F_{op}) in each gene is highly correlated with experimental estimates of protein abundance ($r = 0.59$; figure 2). CAI values are similarly highly correlated with protein abundance ($r = 0.59$).

(b) *The strength of selected codon bias in* Methanococcus maripaludis

To estimate the strength of selected codon bias in the HE genes, we calculated S (the product of the selective difference between codons and the long-term effective population size) as defined previously for Bacteria [6]. The S value of 1.63 for *M. maripaludis* is surprisingly high; it is similar to that for the bacterium *E. coli* ($S = 1.49$), which has a much faster growth rate and more than twice as many rRNA and tRNA genes [6].

The previous S value estimates the strength of selected bias across four pairs of codons for twofold degenerate AAs. Calculating S values for these AAs individually and contrasting *M. maripaludis* and *E. coli* (table 2), we find that S is not significantly different between species (paired t-test, $p = 0.42$). However, analogous S values for fourfold degenerate AAs (see electronic supplementary material) are significantly lower in *M. maripaludis* than in *E. coli* ($p < 0.01$). Comparing between AAs within species, S values are significantly lower ($p = 0.03$) for fourfold than for twofold degenerate AAs in *M. maripaludis*, but not significantly different ($p = 0.36$) in *E. coli*.

4. DISCUSSION

Codon usage varies across the *M. maripaludis* genome, with a single major trend among genes. Various

observations are consistent with selection for translationally optimal codons being the cause of this variation: (i) HE genes lie towards one end of this major trend, (ii) putative translationally optimal codons can be identified for all AAs except Glu (table 1), including codons (UUC, UAC, AUC, AAC) that are optimal in all species owing to their complementarity to the only tRNA anti-codons available [6], and (iii) the frequency of these codons in a gene (F_{op}) is highly correlated with abundance of the encoded protein (figure 2). Strangely, Xia *et al.* [14] reported that CAI values did not correlate well with their protein abundance data; we find the opposite (see electronic supplementary material).

Two additional points arise from the gene expression data. First, data are available for only 967 genes (56% of the total). This may indicate that some predicted open reading frames do not in fact encode proteins, or that many genes were not expressed to a measurable extent under the growth conditions used to estimate protein abundance. Consistent with this, the genes lacking expression level data have lower F_{op} values (median 0.39) than others (median 0.44). Second, codon usage bias is expected to be most strongly selected during periods of exponential growth, and so F_{op} values may correlate less well with expression data (such as those used here) collected under other growth conditions. Interestingly, the HE genes have stronger bias than would be predicted given the overall observed relationship of F_{op} with expression level (figure 2), as expected if these HE genes are even more highly expressed, relative to other genes, during periods at maximum growth rate.

In a recent analysis of selected codon usage bias (S) in the genomes of 214 prokaryotes, the 26 Archaea seemed to conform to the same trends as the 188 Bacteria [8], but here the S value for *M. maripaludis* is high in comparison with bacterial species with a similarly long minimal doubling time (more than 2 h) or with similarly small numbers of rRNA operons and tRNA genes [6,7]. Interestingly, the pattern of bias in *M. maripaludis* differs from that in *E. coli*, the archetypal example of selected codon usage bias in Bacteria (table 2). While the strength of selected bias is similar in the two species for twofold degenerate AAs, the bias is reduced for fourfold degenerate AAs in *M. maripaludis*. For twofold degenerate AAs there is typically only one form of tRNA and one codon is favoured over another because it better matches the anti-codon. For fourfold degenerate AAs there are usually multiple species of tRNA, with abundances largely determined by gene copy number [19]. In *E. coli* (and other Bacteria), strongly selected codon usage bias for fourfold degenerate AAs is associated with increased copy numbers of cognate tRNA genes; *E. coli* K-12 has 26 tRNA genes for these AAs. In contrast, *M. maripaludis* has only 10 genes—one copy for each of two different tRNAs for each of the five AAs; if this explains the weak selected codon usage bias, it remains unclear why tRNA gene duplication has not been favoured.

In conclusion, it is surprising that translational selection in *M. maripaludis* has had a strong impact on codon usage for some (twofold degenerate) but not other (fourfold degenerate) AAs. It will be interesting to extend this study to other species to see whether this is a general difference between Bacteria and Archaea.

We thank Kai Zeng for discussion of the extension of S values to four codon families. L.R.E. was funded by a studentship from the BBSRC.

1 Bulmer, M. 1991 The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129**, 897–907.
2 Sharp, P. M. & Li, H. 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38. (doi:10.1007/BF02099948)
3 Ikemura, T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34.
4 Ehrenberg, M. & Kurland, C. G. 1984 Costs of accuracy determined by a maximal growth rate constraint. *Quart. Rev. Biophys.* **17**, 45–82. (doi:10.1017/S0033583500005254)
5 Rocha, E. P. C. 2004 Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14**, 2279–2286. (doi:10.1101/gr.2896904)
6 Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. 2005 Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**, 1141–1153. (doi:10.1093/nar/gki242)
7 Sharp, P. M., Emery, L. R. & Zeng, K. 2010 Forces that influence the evolution of codon bias. *Phil. Trans. R. Soc. B* **365**, 1203–1212. (doi:10.1098/rstb.2009.0305)
8 Vieira-Silva, S. & Rocha, E. P. C. 2010 The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808. (doi:10.1371/journal.pgen.1000808)
9 Karlin, S., Mrazek, J., Ma, J. & Brocchieri, L. 2005 Predicted highly expressed genes in archaeal genomes. *Proc. Natl Acad. Sci. USA* **102**, 7303–7308. (doi:10.1073/pnas.0502313102)
10 Henry, I. & Sharp, P. M. 2007 Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.* **24**, 10–12. (doi:10.1093/molbev/msl148)
11 Dos Reis, M., Savva, R. & Wernisch, L. 2004 Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044. (doi:10.1093/nar/gkh834)
12 Jones, W. J., Paynter, M. J. B. & Gupta, R. 1983 Characterization of *Methanococcus maripaludis* sp. nov., a new methanogen isolated from salt marsh sediment. *Arch. Microbiol.* **135**, 91–97. (doi:10.1007/BF00408015)
13 Hendrickson, E. L. *et al.* 2004 Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. *J. Bacteriol.* **186**, 6956–6969. (doi:10.1128/JB.186.20.6956-6969.2004)
14 Xia, Q. *et al.* 2006 Quantitative proteomics of the archaeon *Methanococcus maripaludis* validated by microarray analysis and real time PCR. *Mol. Cell. Proteomics* **5**, 868–881. (doi:10.1074/mcp.M500369-MCP200)

15 Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. & Di Paola, G. 1985 ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical design and usage. *Comp. Appl. Biosci.* **1**, 167–172.
16 Chan, P. P. & Lowe, T. M. 2009 GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93–D97. (doi:10.1093/nar/gkn787)
17 Wright, F. 1990 The 'effective number of codons' used in a gene. *Gene* **87**, 23–29. (doi:10.1016/0378-1119(90)90491-9)
18 Sharp, P. M. & Li, H. 1987 The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295. (doi:10.1093/nar/15.3.1281)
19 Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T. 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 143–155. (doi:10.1016/S0378-1119(99)00225-5)